



**HAL**  
open science

# Nonparametric estimation of the stationary density and the transition density of a Markov chain

Claire Lacour

► **To cite this version:**

Claire Lacour. Nonparametric estimation of the stationary density and the transition density of a Markov chain. 2006. hal-00115457v1

**HAL Id: hal-00115457**

**<https://hal.science/hal-00115457v1>**

Preprint submitted on 21 Nov 2006 (v1), last revised 9 Jan 2008 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# NONPARAMETRIC ESTIMATION OF THE STATIONARY DENSITY AND THE TRANSITION DENSITY OF A MARKOV CHAIN

Claire Lacour<sup>1</sup>

April 2005

ABSTRACT. In this paper, we study first the problem of nonparametric estimation of the stationary density  $f$  of a discrete-time Markov chain  $(X_i)$ . We consider a collection of projection estimators on finite dimensional linear spaces. We select an estimator among the collection by minimizing a penalized contrast. The same technique enables to estimate the density  $g$  of  $(X_i, X_{i+1})$  and so to provide an adaptive estimator of the transition density  $\pi = g/f$ . We give bounds in  $L^2$  norm for these estimators and we show that they are adaptive in the minimax sense over a large class of Besov spaces. Some examples and simulations are also provided.

*Keywords:* Adaptive estimation; Markov Chain; Stationary density; Transition density; Model selection; Penalized contrast; Projection estimators.

## 1. INTRODUCTION

Nonparametric estimation is now a very rich branch of statistical theory. The case of i.i.d. observations is the most detailed but many authors are also interested in the case of Markov processes. Early results are stated by Roussas (1969), who studies nonparametric estimators of the stationary density and the transition density of a Markov chain. He considers kernel estimators and assumes that the chain satisfies the strong Doeblin's condition  $(D_0)$  (see Doob (1953) p.221). He shows consistency and asymptotic normality of his estimator. Several authors tried to consider weaker assumptions than the Doeblin's condition. Rosenblatt (1970) introduces an other condition, denoted by  $(G_2)$ , and he gives results on the bias and the variance of the kernel estimator of the invariant density in this weaker framework. Yakowitz (1989) improves also the result of asymptotic normality by considering a Harris-condition.

---

<sup>1</sup>MAP5, Université Paris 5, 45 rue des Saints-Pères 75270 Paris Cedex 06, France.  
email: lacour@math-info.univ-paris5.fr

The study of kernel estimators is completed by Masry and Györfi (1987) who find sharp rates for this kind of estimators of the stationary density and by Basu and Sahoo (1998) who prove a Berry-Esseen inequality under the condition  $(G_2)$  of Rosenblatt. Other authors are interested in the estimation of the invariant distribution and the transition density in the non-stationary case: Doukhan and Ghindès (1983) bound the integrated risks for any initial distribution. In Hernández-Lerma et al. (1988), recursive estimators for a non-stationary Markov chain are described. Liebscher (1992) gives results for the invariant density in this non-stationary framework using a condition denoted by  $(D_1)$  derived from the Doeblin's condition but weaker than  $(D_0)$ . All the above papers deal with kernel estimators. Among those who are not interested in such estimators, let us mention Bosq (1973) who studies an estimator of the stationary density by projection on a Fourier basis, Prakasa Rao (1978) who outlines a new estimator for the stationary density by using delta-sequences and Gillert and Wartenberg (1984) who present estimators based on Hermite bases or trigonometric bases.

The recent work of Cléménçon (1999) allows to measure the performance of all these estimators since he proves lower bounds for the minimax rates and gives thus the optimal convergence rates for the estimation of the stationary density and the transition density. Cléménçon also provides an other kind of estimator for the stationary density and for the transition density, that he obtains by projection on wavelet bases. He presents an adaptive procedure which is "quasi-optimal" in the sense that the procedure reaches almost the optimal rate but with a logarithmic loss. He needs other conditions than those we cited above and in particular a minoration condition derived from Nummelin's (1984) works. In this paper, we will use the same condition.

The aim of this paper is to estimate the stationary density of a discrete-time Markov chain and its transition density. We consider an irreducible positive recurrent Markov chain  $(X_n)$  with a stationary density denoted by  $f$ . We suppose that the initial density is  $f$  (hence the process is stationary) and we construct an estimator  $\tilde{f}$  from the data  $X_1, \dots, X_n$ . Then, we study the mean integrated squared error  $\mathbb{E}\|\tilde{f} - f\|_2^2$  and its convergence rate. The same technique enables to estimate the density  $g$  of  $(X_i, X_{i+1})$  and so to provide an estimator of the transition density  $\pi = g/f$ , called the quotient estimator.

An adaptative procedure is proposed for the two estimations and it is proved that both resulting estimators reach the optimal minimax rates without additive logarithmic factor.

We will use here some technical methods known as the Nummelin splitting technique (see Nummelin (1984), Meyn and Tweedie (1993) or Höpfner and Löcherbach (2003)). This method allows to reduce the general state space Markov chain theory to the countable space theory. Actually, the splitting of the original chain creates an artificial accessible atom and we will use the hitting times to this atom to decompose the chain, as we would have done for a countable space chain.

To build our estimator of  $f$ , we use model selection via penalization as described in Barron et al. (1999). First, estimators by projection denoted by  $\hat{f}_m$  are considered. The index  $m$  denotes the model, i.e. the subspace to which the estimator belongs. Then the model selection technique allows to select automatically an estimator  $\hat{f}_{\hat{m}}$  from the collection of estimators  $(\hat{f}_m)$ . The estimator of  $g$  is built in the same way. The collections of models that we consider here include wavelets but also trigonometric polynomials and piecewise polynomials.

This paper is organized as follows. In section 2, we present our assumptions on the Markov chain and on the collections of models. We give also examples of chains and models. Section 3 is devoted to estimation of the stationary density and contains examples and simulations. In section 4, the estimation of the transition density is explained. The proofs are gathered in the last section, which contains also a presentation of the Nummelin splitting technique.

## 2. THE FRAMEWORK

**2.1. Assumptions on the Markov chain.** We consider an irreducible Markov chain  $(X_n)$  taking its values in the real line  $\mathbb{R}$ . We suppose that  $(X_n)$  is positive recurrent, i.e. it admits a stationary probability measure  $\mu$  (for more details, we refer to Meyn and Tweedie (1993)). We assume that the distribution  $\mu$  has a density  $f$  with respect to the Lebesgue measure and it is this quantity that we want to estimate. Since the number of observations is finite,  $f$  is estimated on a compact set only. Without loss of generality, this compact set is assumed to be equal to  $[0, 1]$  and, from now,  $f$  denotes actually the restriction of the transition density to  $[0, 1]$ . More precisely, the Markov process is supposed to satisfy the following assumptions:

- A1.  $(X_n)$  is irreducible and positive recurrent.
- A2. The distribution of  $X_0$  is equal to  $\mu$ , thus the chain is (strictly) stationary.
- A3. The stationary density  $f$  belongs to  $L^\infty([0, 1])$  i.e.  $\sup_{x \in [0, 1]} |f(x)| < \infty$
- A4. The chain is strongly aperiodic, i.e. it satisfies the following minorization condition: there is some function  $h : [0, 1] \mapsto [0, 1]$  with  $\int h d\mu > 0$  and a positive distribution  $\nu$  such that, for all event  $A$  and for all  $x$ ,

$$P(x, A) \geq h(x)\nu(A)$$

where  $P$  is the transition kernel of  $(X_n)$ .

- A5. The chain is geometrically ergodic, i.e. there exists a function  $V > 0$  finite and a constant  $\rho \in (0, 1)$  such that, for all  $n \geq 1$

$$\|P^n(x, \cdot) - \mu\|_{TV} \leq V(x)\rho^n$$

where  $\|\cdot\|_{TV}$  is the total variation norm.

We can remark that condition A3 implies that  $f$  belongs to  $L^2([0, 1])$  where  $L^2([0, 1]) = \{t : \mathbb{R} \mapsto \mathbb{R}, \text{Supp}(t) \subset [0, 1] \text{ and } \|t\|^2 = \int_0^1 t^2(x)dx < \infty\}$ .

Notice that, if the chain is aperiodic, condition A4 holds, at least for some  $m$ -skeleton (i.e. a chain with transition probability  $P^m$ ) (see Theorem 5.2.2 in Meyn and Tweedie (1993)). This minorization condition is used in the Nummelin splitting technique and is also required in Cléménçon (1999).

The last assumption, which is called geometric regularity by Cléménçon (2000), means that the convergence of the chain to the invariant distribution is geometrically fast. In Meyn and Tweedie (1993), we find a slightly different condition (replacing the total variation norm by the  $V$ -norm). This condition, which is sufficient for A5, is widely used in Monte Carlo Markov Chain literature because it guarantees central limit theorems and enables to simulate laws via a Markov chain (see for example Jarner and Hansen (2000), Roberts and Rosenthal (1998) or Meyn and Tweedie (1994)).

The following subsection gives some examples of Markov chains satisfying hypotheses A1–A5.

## 2.2. Examples of chains.

2.2.1. *Diffusion processes.* We consider the process  $(X_{i\Delta})_{1 \leq i \leq n}$  where  $\Delta > 0$  is the observation step and  $(X_t)_{t \geq 0}$  is defined by

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t$$

where  $W$  is the standard Brownian motion. We suppose that the drift function  $b$  and the diffusion coefficient  $\sigma$  satisfy the following conditions, given by Leblanc (1997):

- (1)  $\exists b, \forall |x|, \quad |b(x)| \leq b(1 + |x|),$
- (2)  $\exists \gamma > 0, \exists R, \forall |x| \geq R, \quad xb(x) \leq -\gamma|x|,$
- (3)  $\exists \sigma_1^2, \exists \sigma_0^2, \forall x, \quad \sigma_1^2 \geq \sigma^2(x) \geq \sigma_0^2 > 0$
- (4)  $\exists L, \forall (x, y), \quad |\sigma(x) - \sigma(y)| \leq L|x - y|^{1/2}.$

Then the discretized process  $(X_{i\Delta})_{1 \leq i \leq n}$  satisfies assumptions A1–A5.

2.2.2. *Nonlinear AR(1) processes.* Let us consider the following process

$$X_n = \varphi(X_{n-1}) + \varepsilon_{X_{n-1}, n}$$

where  $\varepsilon_{x,n}$  has a positive density  $l_x$  with respect to the Lebesgue measure, which does not depend on  $n$ . We suppose that  $\varphi$  is bounded on any compact set and that there exist  $M > 0$  and  $\rho < 1$  such that, for all  $|x| > M$ ,  $|\varphi(x)| < \rho|x|$ . Mokkadem (1987) proves that if there exists  $s > 0$  such that  $\sup_x \mathbb{E}|\varepsilon_{x,n}|^s < \infty$ , then the chain is geometrically ergodic. If we assume furthermore that  $l_x$  has a lower bound then the chain satisfies all the previous assumptions.

2.2.3. *ARX (1,1) models.* The nonlinear process ARX(1,1) is defined by

$$X_n = F(X_{n-1}, Z_n) + \xi_n$$

where  $F$  is bounded and  $(\xi_n), (Z_n)$  are independent sequences of i.i.d. random variables. We suppose that the distribution of  $Z_n$  has a positive density  $l$  with respect to the Lebesgue measure. At last, we assume that there exist positive reals  $a$  and  $\alpha$  such that  $|x| \leq a + \|F\|_\infty \Rightarrow l(x) \geq \alpha$ .

Then the process  $(X_n)$  satisfies assumptions A1–A5 (see Doukhan (1994) p.101).

2.2.4. *ARCH process.* The considered model is

$$X_{n+1} = F(X_n) + G(X_n)\varepsilon_{n+1}$$

where  $F$  and  $G$  are continuous functions and for all  $x$ ,  $G(x) \neq 0$ . We suppose that the distribution of  $\varepsilon_n$  has a positive and continuous density with respect to the Lebesgue measure and that there exists  $s \geq 1$  such that  $\mathbb{E}|\varepsilon_n|^s < \infty$ . The chain  $(X_i)$  satisfies assumptions A1–A5 if (see Doukhan (1994) p.106):

$$\limsup_{|x| \rightarrow \infty} \frac{|F(x)| + |G(x)|(\mathbb{E}|\varepsilon_n|^s)^{1/s}}{|x|} < 1.$$

2.3. **Assumptions on the models.** In order to estimate  $f$ , we need to introduce some collections of models. The assumptions on the models are the following:

M1. Each  $S_m$  is a linear subspace of  $(L^\infty \cap L^2)([0, 1])$  with dimension  $D_m \leq \sqrt{n}$

M2. Let

$$\phi_m = \frac{1}{\sqrt{D_m}} \sup_{t \in S_m \setminus \{0\}} \frac{\|t\|_\infty}{\|t\|}$$

There exists a real  $r_0$  such that for all  $m$ ,  $\phi_m \leq r_0$ .

This assumption ( $L^2$ - $L^\infty$  connexion) is introduced by Barron et al. (1999) and can be written:

$$(1) \quad \forall t \in S_m \quad \|t\|_\infty \leq r_0 \sqrt{D_m} \|t\|.$$

We get then a set of models  $(S_m)_{m \in \mathcal{M}_n}$  where  $\mathcal{M}_n = \{m, D_m \leq \sqrt{n}\}$ . We need now a last assumption regarding the whole collection, which ensures that, for  $m$  and  $m'$  in  $\mathcal{M}_n$ ,  $S_m + S_{m'}$  belongs to the collection of models.

M3. The models are nested, that is for all  $m$ ,  $D_m \leq D_{m'} \Rightarrow S_m \subset S_{m'}$ .

2.4. **Examples of models.** We show here that the assumptions M1-M3 are not too restrictive. Indeed, they are verified for the models spanned by the following bases (see Barron et al. (1999)):

- Histogram basis:  $S_m = \langle \varphi_1, \dots, \varphi_{2^m} \rangle$  with  $\varphi_j = 2^{m/2} \mathbb{1}_{[\frac{j-1}{2^m}, \frac{j}{2^m}[}$  for  $j = 1, \dots, 2^m$ . Here  $D_m = 2^m$ ,  $r_0 = 1$  and  $\mathcal{M}_n = \{1, \dots, \lfloor \ln n / 2 \ln 2 \rfloor\}$  where  $\lfloor x \rfloor$  denotes the floor of  $x$ , i.e. the largest integer less than or equal to  $x$ .
- Trigonometric basis:  $S_m = \langle \varphi_0, \dots, \varphi_{m-1} \rangle$  with  $\varphi_0(x) = 1$ ,  $\varphi_{2j} = \sqrt{2} \cos(2\pi jx) \mathbb{1}_{[0,1]}(x)$ ,  $\varphi_{2j-1} = \sqrt{2} \sin(2\pi jx) \mathbb{1}_{[0,1]}(x)$  for  $j \geq 1$ . For this model  $D_m = m$  and  $r_0 = \sqrt{2}$  hold.

- Regular piecewise polynomial basis:  $S_m$  is spanned by polynomials of degree  $0, \dots, r$  (where  $r$  is fixed) on each interval  $[(j-1)/2^D, j/2^D], j = 1, \dots, 2^D$ . In this case,  $m = (D, r)$ ,  $D_m = (r+1)2^D$  and  $\mathcal{M}_n = \{(D, r), D = 1, \dots, \lfloor \log_2(\sqrt{n}/(r+1)) \rfloor\}$ . We can put  $r_0 = \sqrt{r+1}$ .
- Regular wavelet basis:  $S_m = \langle \psi_{jk}, j = -1, \dots, m, k \in \Lambda(j) \rangle$  where  $\psi_{-1,k}$  points out the translates of the father wavelet and  $\psi_{jk}(x) = 2^{j/2}\psi(2^jx - k)$  where  $\psi$  is the mother wavelet. We assume that the support of the wavelets is included in  $[0, 1]$  and that  $\psi_{-1} = \varphi$  belongs to the Sobolev space  $W_2^r$ . In this framework  $\Lambda(j) = \{0, \dots, K2^j - 1\}$  (for  $j \geq 0$ ) where  $K$  is a constant which depends on the supports of  $\varphi$  and  $\psi$ : for example for the Haar basis  $K = 1$ . We have then  $D_m = \sum_{j=-1}^m |\Lambda(j)| = |\Lambda(-1)| + K(2^{m+1} - 1)$ . Moreover

$$\begin{aligned} \phi_m &\leq \frac{\sum_k |\psi_{-1,k}| + \sum_{j=0}^m 2^{j/2} \sum_k |\psi_{j,k}|}{\sqrt{D_m}} \\ &\leq \frac{\|\varphi\|_\infty \vee \|\psi\|_\infty (1 + \sum_{j=0}^m 2^{j/2})}{\sqrt{(K \wedge |\Lambda(-1)|) 2^{m+1}}} \leq \frac{\|\varphi\|_\infty \vee \|\psi\|_\infty}{K \wedge |\Lambda(-1)|} =: r_0 \end{aligned}$$

### 3. ESTIMATION OF THE STATIONARY DENSITY

3.1. **Decomposition of the risk for the projection estimator.** Let

$$(2) \quad \gamma_n(t) = \frac{1}{n} \sum_{i=1}^n [ \|t\|^2 - 2t(X_i) ].$$

Notice that  $\mathbb{E}(\gamma_n(t)) = \|t - f\|^2 - \|f\|^2$  and therefore  $\gamma_n(t)$  is the empirical version of the  $L^2$  distance between  $t$  and  $f$ . Thus,  $\hat{f}_m$  is defined by

$$(3) \quad \hat{f}_m = \arg \min_{t \in S_m} \gamma_n(t)$$

where  $S_m$  is a subspace of  $L^2$  which satisfies M2. Although this estimator depends on  $n$ , no index  $n$  is mentioned in order to simplify the notations. It is also the case for all the estimators in this paper.

A more explicit formula for  $\hat{f}_m$  is easy to derive:

$$(4) \quad \hat{f}_m = \sum_{\lambda \in \Lambda} \hat{\beta}_\lambda \varphi_\lambda, \quad \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n \varphi_\lambda(X_i)$$

where  $(\varphi_\lambda)_{\lambda \in \Lambda}$  is an orthonormal basis of  $S_m$ . Note that

$$\mathbb{E}(\hat{f}_m) = \sum_{\lambda \in \Lambda} \langle f, \varphi_\lambda \rangle \varphi_\lambda,$$

which is the projection of  $f$  on  $S_m$ .

In order to evaluate the quality of this estimator, we now compute the mean integrated squared error  $\mathbb{E}\|f - \hat{f}_m\|^2$  (often denoted by MISE).

**Proposition 1.** *Let  $X_n$  be a Markov chain which satisfies Assumptions A1–A4 and  $S_m$  be a subspace of  $L^2$  with dimension  $D_m \leq n$ . If  $S_m$  satisfies the condition M2, then the estimator  $\hat{f}_m$  defined by (3) satisfies*

$$\mathbb{E}\|f - \hat{f}_m\|^2 \leq d^2(f, S_m) + C \frac{D_m}{n}$$

where  $C$  is a constant which does not depend on  $n$ .

To compute the bias term  $d(f, S_m)$ , we assume that  $f$  belongs to the Besov space  $B_{2,\infty}^\alpha([0, 1])$ . Let us recall the definition of  $B_{2,\infty}^\alpha([0, 1])$ . Let

$$\Delta_h^r f(x) = \sum_{k=0}^r (-1)^{r-k} \binom{r}{k} f(x + kh)$$

the  $r$ th difference operator with step  $h$  and

$$\omega_r(f, t) = \sup_{|h| \leq t} \|\Delta_h^r f\|$$

the  $r$ th modulus of smoothness of  $f$  where we recall that  $\|t\|^2 = \int_0^1 t^2(x) dx$ . We say that  $f$  is in the Besov space  $B_{2,\infty}^\alpha([0, 1])$  if

$$\sup_{t>0} t^{-\alpha} \omega_r(f, t) < \infty$$

for  $r = \lfloor \alpha \rfloor + 1$  or, equivalently, for  $r$  an integer larger than  $\alpha$ . Notice that when  $\alpha$  is an integer, the Besov space  $B_{2,\infty}^\alpha([0, 1])$  contains the Sobolev space  $W_2^\alpha$  (see DeVore and Lorentz (1993) p.51–55).

Hence, we have the following corollary.

**Corollary 1.** *Let  $X_n$  be a Markov chain which satisfies Assumptions A1–A4. Assume that the stationary density  $f$  belongs to  $B_{2,\infty}^\alpha([0, 1])$  and that  $S_m$  is one of the spaces mentioned in section 2.4 (with the regularity of polynomials and wavelets*

larger than  $\alpha - 1$ ). If we choose  $D_m = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$ , then the estimator defined by (3) satisfies

$$\mathbb{E}\|f - \hat{f}_m\|^2 = O(n^{-\frac{2\alpha}{2\alpha+1}})$$

We can notice that we obtain the same rate than in the i.i.d. case (see Donoho et al. (1996)). Actually, Cléménçon (1999) proves that  $n^{-\frac{2\alpha}{2\alpha+1}}$  is the optimal rate in the minimax sense in the Markovian framework. With very different theoretical tools, Tribouley and Viennet (1998) show that this rate is also reached in the case of the univariate density estimation of  $\beta$ -mixing random variables by using a wavelet estimator. We can remark that our assumption A5 implies the geometrical decreasing of  $\beta$ -mixing coefficients but that, until now, we did not use this assumption.

However, the choice  $D_m = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$  is possible only if we know the regularity  $\alpha$  of the unknown  $f$ . But generally, it is not the case. It is the reason why we construct an adaptive estimator, i.e. an estimator which achieves the optimal rate without requiring the knowledge of  $\alpha$  or any information about  $f$ .

**3.2. Adaptive estimation.** Let  $(S_m)_{m \in \mathcal{M}_n}$  be a collection of models as described in section 2.3. For each  $S_m$ ,  $\hat{f}_m$  is defined as above by (3). Next, we choose  $\hat{m}$  among the family  $\mathcal{M}_n$  such that

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} [\gamma_n(\hat{f}_m) + \text{pen}(m)]$$

where  $\text{pen}$  is a penalty function to be specified later. We denote  $\tilde{f} = \hat{f}_{\hat{m}}$  and we bound the  $L^2$ -risk  $\mathbb{E}\|f - \tilde{f}\|$  as follows.

**Theorem 1.** *Let  $X_n$  be a Markov chain which satisfies Assumptions A1–A5 and  $(S_m)_{m \in \mathcal{M}_n}$  be a collection of models satisfying Assumptions M1–M3. Then the estimator defined by*

$$(5) \quad \tilde{f} = \hat{f}_{\hat{m}} \quad \text{where} \quad \hat{m} = \arg \min_{m \in \mathcal{M}_n} [\gamma_n(\hat{f}_m) + \text{pen}(m)],$$

with

$$(6) \quad \text{pen}(m) = K_0 \frac{D_m}{n}$$

(where  $K_0$  is a constant which depends on the chain) satisfies

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq 3 \inf_{m \in \mathcal{M}_n} \{d^2(f, S_m) + \text{pen}(m)\} + \frac{C_1}{n}$$

where  $C_1$  does not depend on  $n$ .

The constant  $K_0$  in the penalty is equal to  $C \max(r_0^2, 1) \frac{1 + \|f\|_\infty \mathbb{E}_{E_1}(s^{\tau^*})}{(\ln s)^2}$  where  $C$  is a numerical constant. The number  $r_0$  is known and depends on the chosen base (see subsection 2.3). The mention of  $\|f\|_\infty$  in the penalty term seems to be a problem, seeing that  $f$  is unknown. Actually, we could replace  $\|f\|_\infty$  by  $\|\hat{f}\|_\infty$  with  $\hat{f}$  an estimator of  $f$ . This is a bit technical and useless here since there are some terms in  $K_0$  which are not computable either. Indeed,  $E_1$  is an artificial atom created by the Nummelin splitting technique and  $\tau^*$  is the first return time of the split chain in this atom. The number  $s$  is a real larger than 1 such that  $\sup_{x \in E_1} \mathbb{E}_x(s^{\tau^*}) < \infty$  and its existence is guaranteed by the condition of geometric ergodicity A5. Actually, the computation of the penalty is generally "hand-adjusted" and then we do not have to know accurately the number  $s$  or  $\|f\|_\infty$ .

**Corollary 2.** *Let  $X_n$  be a Markov chain which satisfies Assumptions A1–A5 and  $(S_m)_{m \in \mathcal{M}_n}$  be a collection of models mentioned in section 2.4 (with the regularity of polynomials and wavelets larger than  $\alpha - 1$ ). If  $f$  belongs to  $B_{2,\infty}^\alpha([0, 1])$ , with  $\alpha > 1/2$ , then the estimator defined by (5) and (6) satisfies*

$$\mathbb{E}\|\tilde{f} - f\|^2 = O(n^{-\frac{2\alpha}{2\alpha+1}})$$

**Remark 1.** *When  $\alpha > \frac{1}{2}$ ,  $B_{2,\infty}^\alpha([0, 1]) \subset C[0, 1]$  (where  $C[0, 1]$  is the set of the continuous functions with support in  $[0, 1]$ ) and then the assumption A3  $\|f\|_\infty < \infty$  is superfluous.*

We have already noticed that it is the optimal rate in the minimax sense (see the lower bound in Cléménçon (1999)). Note that here the procedure reaches this rate whatever the regularity of  $f$ , without needing to know  $\alpha$ . This result is thus a improvement of the one of Cléménçon (1999), whose adaptive procedure achieves only the rate  $(\log(n)/n)^{\frac{2\alpha}{2\alpha+1}}$ . Moreover, our procedure allows to use more bases (not only wavelets) and is easy to implement.

**3.3. Simulations.** The computation of the previous estimator is very simple. We use the following procedure:

- For each  $m$ , compute  $\gamma_n(\hat{f}_m) + \text{pen}(m)$ . Notice that  $\gamma_n(\hat{f}_m) = -\sum_{\lambda \in \Lambda_m} \hat{\beta}_\lambda^2$  where  $\hat{\beta}_\lambda$  is defined by (4) and is quickly computed.

- Select the argmin  $\hat{m}$  of  $\gamma_n(\hat{f}_m) + \text{pen}(m)$ .
- Choose  $\tilde{f} = \sum_{\lambda \in \Lambda_{\hat{m}}} \hat{\beta}_\lambda \varphi_\lambda$ .

The bases are here adjusted with an affin transform in order to be defined on  $[\min_{1 \leq i \leq n}(X_i), \max_{1 \leq i \leq n}(X_i)]$  instead of  $[0, 1]$ . We consider 3 different bases (see section 2.4): trigonometric basis, histogram basis and piecewise polynomial basis. For the last basis, we use Legendre polynomials with degree 0,1,2 defined by

$$(7) \quad \forall x \in [-1, 1] \quad P_0(x) = \frac{1}{\sqrt{2}}, \quad P_1(x) = \sqrt{\frac{3}{2}}x, \quad P_2(x) = \frac{\sqrt{5}}{2\sqrt{2}}(3x^2 - 1).$$

The variable substitution  $y = 2^{-D-1}(x + 2j - 1)$  allows to use these polynomials on each interval  $[(j-1)/2^D, j/2^D], j = 1, \dots, 2^D, D = 1, \dots, \lfloor \log_2(\sqrt{n}/3) \rfloor$ . We found that a good choice for the penalty function is

$$\text{pen}(m) = 5 \frac{D_m}{n}.$$

We choose to estimate five different distributions:

- the Gamma distribution with scale parameter  $l = 3/2$  and shape parameter  $a = 2$ :  $f(x) = Cx \exp\{-3x/2\} \mathbb{1}_{(0,\infty)}(x)$  with  $C$  such that  $\int f = 1$
- the Exponential distribution with parameter  $l = 0.7$ :  $f(x) = 0.7 \exp\{-0.7x\} \mathbb{1}_{(0,\infty)}(x)$
- the Gaussian distribution with mean 0 and variance 1:  $f(x) = 1/\sqrt{2\pi} \exp\{-x^2/2\}$
- the Beta distribution with parameter  $a = 2$  and  $b = 3$ :  $f(x) = B(1-x)^2 x \mathbb{1}_{(0,1)}(x)$  with  $B$  such that  $\int f = 1$
- the Cauchy distribution:  $f(x) = 1/(\pi(1+x^2))$ .

The chains are simulated with a Metropolis-Hastings algorithm (see for example Gilks et al. (1996)). We choose here  $q(x, y) = 1/\sigma\sqrt{2\pi} \exp\{-(y-x)^2/2\sigma^2\}$  as a candidate transition density, with  $\sigma = 0.2$  for the Beta distribution and  $\sigma = 1$  otherwise. Indeed, for the sake of realism, the support of the candidate distribution must be close to the support of the simulated distribution.

Figures 1–3 illustrate the performance of the method and Table 1 shows the  $L^2$ -risk for different values of  $n$ .

We can compare results of Table 1 with those of Dalelane (2005) who gives results of simulations for i.i.d. random variables. For the density estimation, she uses three types of kernel: Gauss kernel, sinc-kernel (where  $\text{sinc}(x) = \sin(x)/x$ ) and her Cross Validation optimal kernel (denoted by Dal). Table 2 gives her results for the

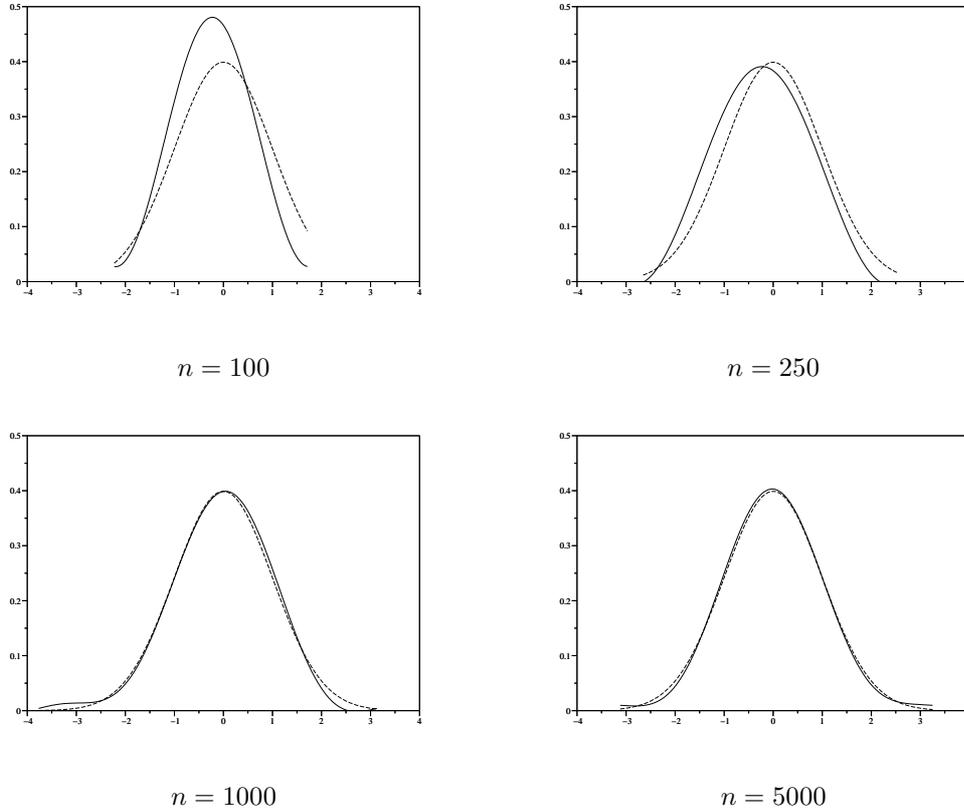


FIGURE 1. Estimator (solid line) and true function (dotted line) for a Gaussian distribution estimated with a trigonometric basis.

Gaussian density and the Gamma distribution with the same parameters that we used (2 and  $3/2$ ). If we compare the results that she obtains with her optimal kernel and our results with trigonometric or polynomial bases, we observe that her risks are about 5 times less than ours. However this kernel is particularly effective and if we consider the classical kernels, we notice that the results are completely comparable, with a reasonable price for dependency.

So the results are roughly good but we can not pretend that a basis among the others gives better results. We can then imagine a mixed strategy, i.e. a procedure which uses several kinds of bases and which can choose the best basis or, for instance, the best degree for a polynomial basis. These techniques are successfully used in a regression framework by Comte and Rozenholc (2002) or Comte and Rozenholc (2004).

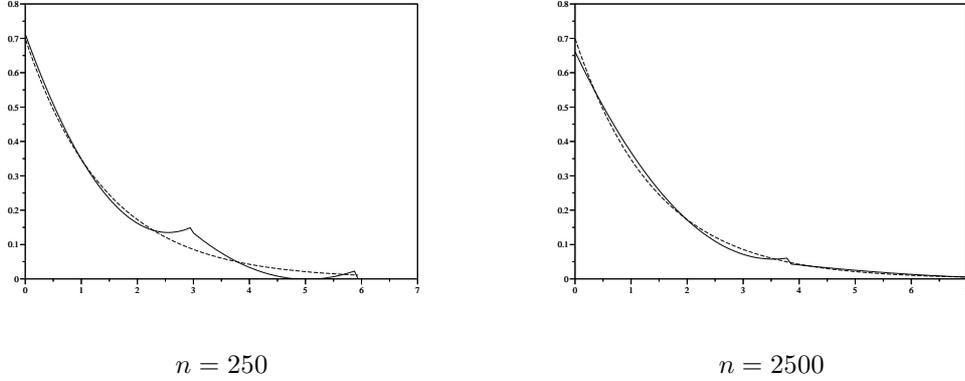


FIGURE 2. Estimator (solid line) and true function (dotted line) for an Exponential distribution estimated with a piecewise polynomial basis.

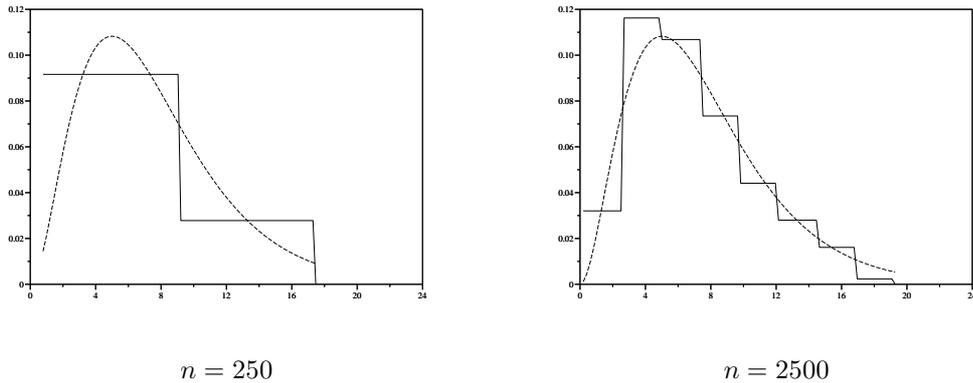


FIGURE 3. Estimator (solid line) and true function (dotted line) for a Gamma distribution estimated with an histogram basis.

We can also notice in Table 1 that some distributions are better estimated than others. An explanation can be found in Mengersen and Tweedie (1996) where results for Metropolis chains, i.e. chains simulated by the Metropolis algorithm, are given. Thus, we know that our simulated chains are irreducible (see Lemma 1.1 in Mengersen and Tweedie (1996)) and satisfy condition A4 with  $\nu = \mu$  and  $h(x) = \varepsilon/d\mathbb{1}_C(x)$  where  $\varepsilon = \inf_{x,y \in C} q(x,y)$  and  $d = \sup_{x \in C} f(x)$  for any compact set  $C$  (see Lemma 2.1 in Mengersen and Tweedie (1996)). They prove next that (for a distribution on the whole real line) the Metropolis chain is geometrically ergodic

| law \ $n$ | 100    | 250    | 500    | 1000   | 2500   | 5000   | basis |
|-----------|--------|--------|--------|--------|--------|--------|-------|
| Gamma     | 0.0626 | 0.0386 | 0.0308 | 0.0211 | 0.0112 | 0.0077 | T     |
|           | 0.0864 | 0.0820 | 0.0703 | 0.0626 | 0.0514 | 0.0283 | H     |
|           | 0.0491 | 0.0237 | 0.0218 | 0.0209 | 0.0092 | 0.0063 | P     |
| Exp       | 0.1284 | 0.0985 | 0.0800 | 0.0680 | 0.0580 | 0.0544 | T     |
|           | 0.0778 | 0.0583 | 0.0419 | 0.0278 | 0.0193 | 0.0112 | H     |
|           | 0.1409 | 0.0241 | 0.0122 | 0.0067 | 0.0036 | 0.0031 | P     |
| Gaussian  | 0.0359 | 0.0102 | 0.0054 | 0.0030 | 0.0012 | 0.0005 | T     |
|           | 0.0571 | 0.0326 | 0.0229 | 0.0170 | 0.0082 | 0.0071 | H     |
|           | 0.0350 | 0.0128 | 0.0074 | 0.0045 | 0.0030 | 0.0027 | P     |
| Beta      | 0.1807 | 0.0732 | 0.0410 | 0.0244 | 0.0111 | 0.0066 | T     |
|           | 0.2100 | 0.1170 | 0.0724 | 0.0444 | 0.0238 | 0.0144 | H     |
|           | 0.4736 | 0.3053 | 0.2385 | 0.2101 | 0.1863 | 0.1803 | P     |
| Cauchy    | 0.0469 | 0.0296 | 0.0191 | 0.0130 | 0.0070 | 0.0040 | T     |
|           | 0.0592 | 0.0522 | 0.0437 | 0.0304 | 0.0182 | 0.0148 | H     |
|           | 0.0446 | 0.0299 | 0.0336 | 0.0262 | 0.0161 | 0.0102 | P     |

TABLE 1. MISE for simulated data with  $\text{pen}(m) = 5D_m/n$ , averaged over  $N = 200$  samples. T: trigonometric basis, H: histogram basis, P: piecewise polynomial basis detailed in (7).

if the tail of  $f$  is at least exponentially decreasing, but if the tail is only polynomial, it is not geometrically ergodic. This may explain for instance why errors are smaller for the Gaussian distribution than for the Cauchy distribution.

#### 4. ESTIMATION OF THE TRANSITION DENSITY

We now suppose that the transition kernel  $P$  has a density  $\pi$ . In order to estimate  $\pi$ , we remark that  $\pi$  can be written  $g/f$  where  $g$  is the density of  $(X_i, X_{i+1})$ . Thus we begin with the estimation of  $g$ . As previously,  $g$  and  $\pi$  are estimated on a compact set which is assumed to be equal to  $[0, 1]^2$ , without loss of generality.

**4.1. Estimation of  $g$ .** We need now a new assumption.

A3'.  $\pi$  belongs to  $L^\infty([0, 1]^2)$ .

| law \ n  | 100    | 500    | 1000   | kernel |
|----------|--------|--------|--------|--------|
| Gamma    | 0.0148 | 0.0052 | 0.0027 | Dal    |
|          | 0.0209 | 0.0061 | 0.0031 | Gauss  |
|          | 0.0403 | 0.0166 | 0.0037 | sinc   |
| Gaussian | 0.0065 | 0.0013 | 0.0008 | Dal    |
|          | 0.0127 | 0.0028 | 0.0016 | Gauss  |
|          | 0.0114 | 0.0026 | 0.0010 | sinc   |

TABLE 2. MISE obtained by Dalelane (2005) for i.i.d. data, averaged over 50 samples

Notice that A3' implies A3. We consider now the following subspaces.

$$S_m^{(2)} = \{t \in L^2([0, 1]^2), \quad t(x, y) = \sum_{\lambda, \mu \in \Lambda_m} \alpha_{\lambda, \mu} \varphi_\lambda(x) \varphi_\mu(y)\}$$

where  $(\varphi_\lambda)_{\lambda \in \Lambda_m}$  is an orthonormal basis of  $S_m$ . Notice that, if we set

$$\phi_m^{(2)} = \frac{1}{D_m} \sup_{t \in S_m^{(2)} \setminus \{0\}} \frac{\|t\|_\infty}{\|t\|},$$

hypothesis M2 implies that  $\phi_m^{(2)}$  is bounded by  $r_0^2$ . The condition M1 must be replaced by the following condition:

M1'. Each  $S_m^{(2)}$  is a linear subspace of  $(L^\infty \cap L^2)([0, 1]^2)$  with dimension  $D_m^2 \leq \sqrt{n}$ .

Let now

$$\gamma_n^{(2)}(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} \{\|t\|^2 - 2t(X_i, X_{i+1})\}.$$

We define as above

$$\hat{g}_m = \arg \min_{t \in S_m^{(2)}} \gamma_n^{(2)}(t)$$

and  $\hat{m}^{(2)} = \arg \min_{m \in \mathcal{M}_n} [\gamma_n^{(2)}(\hat{g}_m) + \text{pen}^{(2)}(m)]$  where  $\text{pen}^{(2)}(m)$  is a penalty function which would be specified later. Lastly, we set  $\tilde{g} = \hat{g}_{\hat{m}^{(2)}}$ .

**Theorem 2.** *Let  $X_n$  be a Markov chain which satisfies Assumptions A1-A2-A3'-A4-A5 and  $(S_m)_{m \in \mathcal{M}_n}$  be a collection of models satisfying Assumptions M1'-M2-M3.*

Then the estimator defined by

$$(8) \quad \tilde{g} = \hat{g}_{\hat{m}^{(2)}} \quad \text{where} \quad \hat{m}^{(2)} = \arg \min_{m \in \mathcal{M}_n} [\gamma_n^{(2)}(\hat{g}_m) + \text{pen}^{(2)}(m)],$$

with

$$(9) \quad \text{pen}^{(2)}(m) = K_0^{(2)} \frac{D_m^2}{n}$$

(where  $K_0^{(2)}$  is a constant which depends on the chain) satisfies

$$\mathbb{E} \|\tilde{g} - g\|^2 \leq 3 \inf_{m \in \mathcal{M}_n} \{d^2(g, S_m^{(2)}) + \text{pen}^{(2)}(m)\} + \frac{C_1}{n}$$

where  $C_1$  does not depend on  $n$ .

The constant  $K_0^{(2)}$  in the penalty is similar to the constant  $K_0$  in Theorem 1 (replacing  $r_0$  by  $r_0^2$  and  $\|f\|_\infty$  by  $\|g\|_\infty$ ). It is compounded of unknown terms but, as previously explained, it is not a problem for practical purposes.

**Corollary 3.** *Let  $X_n$  be a Markov chain which satisfies assumptions A1-A2-A3'-A4-A5 and  $(S_m)_{m \in \mathcal{M}_n}$  be a collection of models mentioned in section 2.4 (with the regularity of polynomials and wavelets larger than  $\alpha-1$ ). If  $g$  belongs to  $B_{2,\infty}^\alpha([0, 1]^2)$ , with  $\alpha > 1$ , then*

$$\mathbb{E} \|\tilde{g} - g\|^2 = O(n^{-\frac{2\alpha}{2\alpha+2}})$$

This rate of convergence is good since it is the minimax rate for density estimation in dimension 2 in the case of i.i.d. random variables (see for instance Ibragimov and Has'minskiĭ (1980)). Let us now proceed to the estimation of the transition density.

4.2. **Estimation of  $\pi$ .** The estimator of  $\pi$  is defined in the following way. Let

$$\tilde{\pi}(x, y) = \begin{cases} \frac{\tilde{g}(x, y)}{\tilde{f}(x)} & \text{if } |\tilde{g}(x, y)| \leq a_n |\tilde{f}(x)| \\ 0 & \text{else} \end{cases}$$

with  $a_n = n^\beta$  and  $\beta < 1/8$ .

We introduce a new assumption:

A6. There exists a positive  $\chi$  such that  $\forall x \in [0, 1], \quad f(x) \geq \chi$ .

**Theorem 3.** *Let  $X_n$  be a Markov chain which satisfies Assumptions A1-A2-A3'-A4-A5-A6 and  $(S_m)_{m \in \mathcal{M}_n}$  be a collection of models mentioned in section 2.4 (with the regularity of polynomials and wavelets larger than  $\alpha - 1$ ). We suppose that the dimension  $D_m$  of the models is such that*

$$\forall m \in \mathcal{M}_n \quad \ln n \leq D_m \leq n^{1/4}.$$

*If  $f$  belongs to  $B_{2,\infty}^\alpha([0,1])$ , with  $\alpha > 1/2$ , then for  $n$  large enough*

- *there exists  $C_1$  and  $C_2$  such that*

$$\mathbb{E}\|\pi - \tilde{\pi}\|^2 \leq C_1 \mathbb{E}\|g - \tilde{g}\|^2 + C_2 \mathbb{E}\|f - \tilde{f}\|^2 + o\left(\frac{1}{n}\right)$$

- *if furthermore  $g$  belongs to  $B_{2,\infty}^\beta([0,1]^2)$  (with  $\beta > 1$ ), then*

$$\mathbb{E}\|\pi - \tilde{\pi}\|^2 = O\left(\sup\left(n^{-\frac{2\beta}{2\beta+2}}, n^{-\frac{2\alpha}{2\alpha+1}}\right)\right)$$

Cléménçon (2000) proved that  $n^{-2\beta/(2\beta+2)}$  is the minimax rate for  $f$  and  $g$  of same regularity  $\beta$ . Notice that in this case the procedure is adaptive and there is no logarithmic loss in the estimation rate contrary to the result of Cléménçon (2000).

If  $g$  belongs to  $B_{2,\infty}^\beta(\mathbb{R}^2)$  (that is to say that we consider the regularity of  $g$  on its whole support and not only on the compact of the observations) then equality  $f(y) = \int g(x,y)dx$  yields that  $f$  belongs to  $B_{2,\infty}^\beta(\mathbb{R})$  and then  $\mathbb{E}\|\pi - \tilde{\pi}\|^2 = O(n^{-\frac{2\beta}{2\beta+2}})$ . Moreover the same rate is achieved if  $\pi$  belongs to  $B_{2,\infty}^\beta(\mathbb{R}^2)$  with  $\beta > 1$ . Indeed formula  $f(y) = \int f(x)\pi(x,y)dx$  implies that  $f$  belongs to  $B_{2,\infty}^\beta(\mathbb{R})$ . Then, by using properties of Besov spaces (see Runst and Sickel (1996) p.192),  $g = f\pi$  belongs to  $B_{2,\infty}^\beta(\mathbb{R}^2)$ .

But it should be remembered that we consider only the restriction of  $f$  or  $\pi$  since the observations are in a compact set. And, as the example of Cléménçon (2000) proves it, the restriction of the stationary density to  $[0,1]$  may be less regular than the restriction of the transition density. The previous procedure has thus the disadvantage that the resulting rate does not depend only on the regularity of  $\pi$  but also on the one of  $f$ .

## 5. PROOFS

**5.1. The Nummelin splitting technique.** This whole subsection is summarized from Höpfner and Löcherbach (2003) p.60–63 and is detailed for the sake of completeness.

The interest of the Nummelin splitting technique is to create a two-dimensional chain (the "split chain"), which contains automatically an atom. Let us recall the definition of an atom. Let  $A$  be a set such that  $\psi(A) > 0$  where  $\psi$  is an irreducibility measure. The set  $A$  is called an atom for the chain  $(X_n)$  with transition kernel  $P$  if there exists a measure  $\nu$  such that  $P(x, B) = \nu(B)$ , for all  $x$  in  $A$  and for all event  $B$ .

Let us now describe the splitting method. Let  $E = [0, 1]$  the state space and  $\mathcal{E}$  the associated  $\sigma$ -field. Each point  $x$  in  $E$  is splitted in  $x_0 = (x, 0) \in E_0 = E \times \{0\}$  and  $x_1 = (x, 1) \in E_1 = E \times \{1\}$ . Each set  $A$  in  $\mathcal{E}$  is splitted in  $A_0 = A \times \{0\}$  and  $A_1 = A \times \{1\}$ . Thus, we have defined a new probability space  $(E^*, \mathcal{E}^*)$  where  $E^* := E_0 \cup E_1$  and  $\mathcal{E}^* = \sigma(A_0, A_1 : A \in \mathcal{E})$ . Using  $h$  defined in A4, a measure  $\lambda$  on  $(E, \mathcal{E})$  splits according to

$$\begin{cases} \lambda^*(A_1) &= \int \mathbf{1}_A(x)h(x)\lambda(dx) \\ \lambda^*(A_0) &= \int \mathbf{1}_A(x)(1-h)(x)\lambda(dx) \end{cases}$$

Notice that  $\lambda^*(A_0 \cup A_1) = \lambda(A)$ . Now the aim is to define a new transition probability  $P^*(\cdot, \cdot)$  on  $(E^*, \mathcal{E}^*)$  to replace the transition kernel  $P$  of  $(X_n)$ . Let

$$P^*(x_i, \cdot) = \begin{cases} \frac{1}{1-h(x)}(P - h \otimes \nu)^*(x, \cdot) & \text{if } i = 0 \text{ and } h(x) > 1 \\ \nu^* & \text{else} \end{cases}$$

where  $\nu$  is the measure introduced in A4 and  $h \otimes \nu$  is a kernel defined by  $h \otimes \nu(x, dy) = h(x)\nu(dy)$ . Consider now a chain  $(X_n^*)$  on  $(E^*, \mathcal{E}^*)$  with one-step transition  $P^*$  and with starting law  $\mu^*$ . The split chain  $(X_n^*)$  has the following properties:

P1. For all  $(A_p)_{0 \leq p \leq N} \in \mathcal{E}^N$  and for all measure  $\lambda$

$$P_\lambda(X_p \in A_p, 0 \leq p \leq N) = P_{\lambda^*}(X_p^* \in A_p \times \{0, 1\}, 0 \leq p \leq N).$$

P2. The split chain is irreducible positive recurrent with stationary distribution  $\mu^*$ .

P3. The set  $E_1$  is an atom for  $(X_n^*)$ .

We can also extend functions  $g : E \mapsto \mathbb{R}$  to  $E^*$  via

$$g^*(x_0) = g(x) = g^*(x_1)$$

Then, the property P1 can be written: for all function  $\mathcal{E}$ -measurable  $g : E^N \mapsto \mathbb{R}$

$$\mathbb{E}_\lambda(g(X_1, \dots, X_N)) = \mathbb{E}_{\lambda^*}(g^*(X_1^*, \dots, X_N^*)).$$

We can say that  $(X_n)$  is a marginal chain of  $(X_n^*)$ . When necessary, the following proofs are decomposed in two steps: first, we assume that the Markov chain has an atom, next we extend the result to the general chain by introducing the artificial atom  $E_1$ .

**5.2. Proof of Proposition 1.** *First step:* We suppose that  $(X_n)$  has an atom  $A$ .

Let  $f_m$  be the orthogonal projection of  $f$  on  $S_m$ . Pythagoras theorem gives us:

$$\mathbb{E}\|f - \hat{f}_m\|^2 = d^2(f, S_m) + \mathbb{E}\|f_m - \hat{f}_m\|^2.$$

We recognize in the right member a bias term and a variance term. According to the expression (4) of  $\hat{f}_m$  the variance term can be written:

$$(10) \quad \mathbb{E}\|f_m - \hat{f}_m\|^2 = \sum_{\lambda \in \Lambda_m} \text{Var}(\hat{\beta}_\lambda) = \sum_{\lambda \in \Lambda_m} \mathbb{E}(\nu_n^2(\varphi_\lambda))$$

where  $\nu_n(t) = \frac{1}{n} \sum_{i=1}^n [t(X_i) - \langle t, f \rangle]$ . By denoting  $\tau = \tau(1) = \inf\{n \geq 1, X_n \in A\}$  and  $\tau(j) = \inf\{n > \tau(j-1), X_n \in A\}$  for  $j \geq 2$ , we can decompose  $\nu_n(t)$  in the following way (see Cléménçon (2001)):

$$(11) \quad \nu_n(t) = \nu_n^{(1)}(t) + \nu_n^{(2)}(t) + \nu_n^{(3)}(t) + \nu_n^{(4)}(t)$$

$$\begin{aligned} \text{with } \nu_n^{(1)}(t) &= \nu_n(t) \mathbf{1}_{\tau > n}, \\ \nu_n^{(2)}(t) &= \frac{1}{n} \sum_{i=1}^{\tau} [t(X_i) - \langle t, f \rangle] \mathbf{1}_{\tau \leq n}, \\ \nu_n^{(3)}(t) &= \frac{1}{n} \sum_{i=1+\tau(1)}^{\tau(l_n)} [t(X_i) - \langle t, f \rangle] \mathbf{1}_{\tau \leq n}, \\ \nu_n^{(4)}(t) &= \frac{1}{n} \sum_{i=\tau(l_n)+1}^n [t(X_i) - \langle t, f \rangle] \mathbf{1}_{\tau \leq n}, \end{aligned}$$

and  $l_n = \sum_{i=1}^n \mathbf{1}_A(X_i)$  (number of visits to the atom  $A$ ). Hence,

$$\nu_n(t)^2 \leq 4\{\nu_n^{(1)}(t)^2 + \nu_n^{(2)}(t)^2 + \nu_n^{(3)}(t)^2 + \nu_n^{(4)}(t)^2\}.$$

- To bound  $\nu_n^{(1)}(t)^2$ , notice that  $|\nu_n(t)| \leq 2\|t\|_\infty$ . And then, by using M2 and (1),  $|\nu_n^{(1)}(t)| \leq 2r_0\sqrt{D_m}\|t\|\mathbf{1}_{\tau>n} \leq 2r_0\sqrt{n}\|t\|\mathbf{1}_{\tau>n}$ . Thus,

$$\mathbb{E}(\nu_n^{(1)}(t)^2) \leq 4r_0^2\|t\|^2nP(\tau > n) \leq 4r_0^2\|t\|^2\frac{\mathbb{E}(\tau^2)}{n}.$$

- We bound the second term in the same way. Since  $|\nu_n^{(2)}(t)| \leq 2(\tau/n)\|t\|_\infty$ , we obtain  $|\nu_n^{(2)}(t)| \leq 2\|t\|r_0(\tau/\sqrt{n})$  and then

$$\mathbb{E}(\nu_n^{(2)}(t)^2) \leq 4r_0^2\|t\|^2\frac{\mathbb{E}(\tau^2)}{n}.$$

- Let us study now the fourth term. As

$$|\nu_n^{(4)}(t)| \leq 2\frac{n - \tau(l_n)}{n}\|t\|_\infty\mathbf{1}_{\tau \leq n} \leq 2\frac{n - \tau(l_n)}{\sqrt{n}}r_0\|t\|\mathbf{1}_{\tau \leq n},$$

we get  $\mathbb{E}(\nu_n^{(4)}(t)^2) \leq 4\frac{r_0^2}{n}\|t\|^2\mathbb{E}((n - \tau(l_n))^2\mathbf{1}_{\tau \leq n})$ .

It remains to bound  $\mathbb{E}((n - \tau(l_n))^2\mathbf{1}_{\tau \leq n})$ :

$$\begin{aligned} \mathbb{E}_\mu((n - \tau(l_n))^2\mathbf{1}_{\tau \leq n}) &= \sum_{k=1}^n \mathbb{E}_\mu((n - k)^2\mathbf{1}_{\tau(l_n)=k}\mathbf{1}_{\tau \leq n}) \\ &= \sum_{k=1}^n (n - k)^2 P_\mu(X_{k+1} \notin A, \dots, X_n \notin A | X_k \in A) P_\mu(X_k \in A) \\ &= \sum_{k=1}^n (n - k)^2 P_A(X_1 \notin A, \dots, X_{n-k} \notin A) \mu(A) \end{aligned}$$

by using the stationarity of  $X$  and the Markov property. Hence

$$\begin{aligned} \mathbb{E}_\mu((n - \tau(l_n))^2\mathbf{1}_{\tau \leq n}) &= \sum_{k=1}^n (n - k)^2 P_A(\tau > n - k) \mu(A) \\ &\leq \sum_{k=1}^{n-1} \frac{\mathbb{E}_A(\tau^4)}{(n - k)^2} \mu(A). \end{aligned}$$

Therefore  $\mathbb{E}_\mu((n - \tau(l_n))^2\mathbf{1}_{\tau \leq n}) = 2\mathbb{E}_A(\tau^4)\mu(A)$ . Finally

$$\mathbb{E}(\nu_n^{(4)}(t)^2) \leq 8r_0^2\|t\|^2\frac{\mu(A)\mathbb{E}_A(\tau^4)}{n}$$

and we can summarize the last three results by

$$(12) \quad \mathbb{E}(\nu_n^{(1)}(t)^2 + \nu_n^{(2)}(t)^2 + \nu_n^{(4)}(t)^2) \leq 8r_0^2\|t\|^2\frac{\mathbb{E}_\mu(\tau^2) + \mu(A)\mathbb{E}_A(\tau^4)}{n}.$$

In particular, if  $t = \varphi_\lambda$ :

$$\mathbb{E}(\nu_n^{(1)}(\varphi_\lambda)^2 + \nu_n^{(2)}(\varphi_\lambda)^2 + \nu_n^{(4)}(\varphi_\lambda)^2) \leq 8r_0^2 \frac{\mathbb{E}_\mu(\tau^2) + \mu(A)\mathbb{E}_A(\tau^4)}{n}.$$

• Last

$$\nu_n^{(3)}(t) \leq \frac{1}{n} \sum_{i=1+\tau(1)}^{\tau(l_n)} [t(X_i) - \langle t, f \rangle].$$

Let us write  $\nu_n^{(3)}(t) \leq \frac{1}{n} \sum_{j=1}^{l_n-1} S_j(t)$  where

$$(13) \quad S_j(t) = \sum_{i=1+\tau(j)}^{\tau(j+1)} (t(X_i) - \langle t, f \rangle).$$

We remark that, according to the Markov property, the  $S_j(t)$  are independent identically distributed and centered. Thus,

$$\mathbb{E}(\nu_n^{(3)}(\varphi_\lambda)^2) \leq \frac{1}{n^2} \sum_{j=1}^n \mathbb{E}|S_j(\varphi_\lambda)|^2.$$

Then, we use Lemma 1 below to bound the expectation of  $\nu_n^{(3)}(\varphi_\lambda)^2$  :

**Lemma 1.** *For all  $m \geq 2$ ,  $\mathbb{E}_\mu|S_j(t)|^m \leq (2\|t\|_\infty)^{m-2} \|f\|_\infty \|t\|^2 \mathbb{E}_A(\tau^m)$ .*

We can then give the bound

$$\mathbb{E}(\nu_n^{(3)}(\varphi_\lambda)^2) \leq \frac{1}{n^2} \sum_{j=1}^n \|f\|_\infty \|\varphi_\lambda\|^2 \mathbb{E}_A(\tau^2) \leq \frac{\|f\|_\infty \mathbb{E}_A(\tau^2)}{n}.$$

Finally

$$\mathbb{E}(\nu_n^2(\varphi_\lambda)) \leq \frac{4}{n} [8r_0^2(\mathbb{E}_\mu(\tau^2) + \mu(A)\mathbb{E}_A(\tau^4)) + \|f\|_\infty \mathbb{E}_A(\tau^2)].$$

Letting  $C = 4[8r_0^2(\mathbb{E}_\mu(\tau^2) + \mu(A)\mathbb{E}_A(\tau^4)) + \|f\|_\infty \mathbb{E}_A(\tau^2)]$ , we obtain with (10)

$$\mathbb{E}\|f_m - \hat{f}_m\|^2 \leq C \frac{D_m}{n}.$$

*Second step:* We do not suppose any more that  $(X_n)$  has an atom.

Let us apply the Nummelin splitting technique to the chain  $(X_n)$  and let

$$(14) \quad \gamma_n^*(t) = \frac{1}{n} \sum_{i=1}^n [ \|t\|^2 - 2t^*(X_i^*) ].$$

We define also

$$(15) \quad \hat{f}_m^* = \arg \min_{t \in S_m} \gamma_n^*(t).$$

Then the property P1 in section 5.1 yields

$$\mathbb{E}\|f - \hat{f}_m^*\|^2 = \mathbb{E}\|f - \hat{f}_m\|^2$$

The split chain having an atom (property P3), we can use the first step to deduce  $\mathbb{E}\|f - \hat{f}_m^*\|^2 = d^2(f, S_m) + CD_m/n$ . It follows that

$$\mathbb{E}\|f - \hat{f}_m\|^2 \leq d^2(f, S_m) + CD_m/n.$$

□

**Proof of Lemma 1:** For all  $j$ ,  $\mathbb{E}_\mu|S_j(t)|^m = \mathbb{E}_\mu|S_1(t)|^m = \mathbb{E}_\mu|\sum_{i=\tau+1}^{\tau(2)} \bar{t}(X_i)|^m$  where  $\bar{t} = t - \langle t, f \rangle$ . Thus

$$\begin{aligned} \mathbb{E}_\mu|S_j(t)|^m &= \sum_{k < l} \mathbb{E} \left( \left| \sum_{i=k+1}^l \bar{t}(X_i) \right|^m \middle| \tau = k, \tau(2) = l \right) P(\tau = k, \tau(2) = l) \\ &\leq \sum_{k < l} (2\|t\|_\infty(l-k))^{m-2} \mathbb{E} \left( \left| \sum_{i=k+1}^l \bar{t}(X_i) \right|^2 \middle| \tau = k, \tau(2) = l \right) P(\tau = k, \tau(2) = l) \\ &\leq \sum_{k < l} (2\|t\|_\infty(l-k))^{m-2} \mathbb{E} \left( \left| \sum_{i=k+1}^l \bar{t}(X_1) \right|^2 \middle| \tau = k, \tau(2) = l \right) P(\tau = k, \tau(2) = l) \end{aligned}$$

since, under  $\mu$ , the  $X_i$  have the same distribution.

$$\begin{aligned} \mathbb{E}_\mu|S_j(t)|^m &\leq \sum_{k < l} (2\|t\|_\infty)^{m-2} (l-k)^m \mathbb{E}(t^2(X_1)) P(\tau = k, \tau(2) = l) \\ &\leq \sum_{k < l} (2\|t\|_\infty)^{m-2} (l-k)^m \|f\|_\infty \|t\|^2 P(\tau = k, \tau(2) = l) \\ &\leq (2\|t\|_\infty)^{m-2} \mathbb{E}(|\tau(2) - \tau|^m) \|f\|_\infty \|t\|^2. \end{aligned}$$

We conclude by using the Markov property. □

**5.3. Proof of Corollary 1.** According to Proposition 1

$$\mathbb{E}\|f - \hat{f}_m\|^2 \leq d^2(f, S_m) + C \frac{D_m}{n}.$$

Then we use Lemma 12 in Barron et al. (1999) which ensures that (for piecewise polynomials or wavelets having a regularity larger than  $\alpha - 1$  and for trigonometric polynomials)

$$d^2(f, S_m) = O(D_m^{-2\alpha})$$

Thus,

$$\mathbb{E}\|f - \hat{f}_m\|^2 = O(D_m^{-2\alpha} + \frac{D_m}{n})$$

In particular, if  $D_m = \lfloor n^{\frac{1}{1+2\alpha}} \rfloor$ , then  $\mathbb{E}\|f - \hat{f}_m\|^2 = O(n^{-\frac{2\alpha}{1+2\alpha}})$ .  $\square$

**5.4. Proof of Theorem 1.** *First step:* We suppose that  $(X_n)$  has an atom  $A$ .

Let  $m$  in  $\mathcal{M}_n$ . The definition of  $\hat{m}$  yields that

$$\gamma_n(\hat{f}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma_n(f_m) + \text{pen}(m).$$

This leads to

$$(16) \quad \|\hat{f}_{\hat{m}} - f\|^2 \leq \|f_m - f\|^2 + 2\nu_n(\hat{f}_{\hat{m}} - f_m) + \text{pen}(m) - \text{pen}(\hat{m})$$

where  $\nu_n(t) = \frac{1}{n} \sum_{i=1}^n [t(X_i) - \langle t, f \rangle]$ .

**Remark 2.** *If  $t$  is deterministic,  $\nu_n(t)$  can actually be written  $\nu_n(t) = \frac{1}{n} \sum_{i=1}^n [t(X_i) - \mathbb{E}(t(X_i))]$ .*

We set  $B(m') = \{t \in S_m + S_{m'}, \|t\| = 1\}$ . Let us write now

$$\begin{aligned} 2\nu_n(\hat{f}_{\hat{m}} - f_m) &= 2\|\hat{f}_{\hat{m}} - f_m\| \nu_n\left(\frac{\hat{f}_{\hat{m}} - f_m}{\|\hat{f}_{\hat{m}} - f_m\|}\right) \\ &\leq 2\|\hat{f}_{\hat{m}} - f_m\| \sup_{t \in B(\hat{m})} \nu_n(t) \leq \frac{1}{5}\|\hat{f}_{\hat{m}} - f_m\|^2 + 5 \sup_{t \in B(\hat{m})} \nu_n(t)^2 \end{aligned}$$

by using inequality  $2xy \leq \frac{1}{5}x^2 + 5y^2$ . Thus,

$$(17) \quad 2\mathbb{E}|\nu_n(\hat{f}_{\hat{m}} - f_m)| \leq \frac{1}{5}\mathbb{E}\|\hat{f}_{\hat{m}} - f_m\|^2 + 5\mathbb{E}\left(\sup_{t \in B(\hat{m})} \nu_n(t)^2\right).$$

Consider decomposition (11) of  $\nu_n(t)$  again. We can write

$$\sup_{t \in B(\hat{m})} \nu_n^{(3)}(t)^2 \leq p(m, \hat{m}) + \sum_{m' \in \mathcal{M}_n} \left[ \sup_{t \in B(m')} \nu_n^{(3)}(t)^2 - p(m, m') \right]$$

where  $p(\cdot, \cdot)$  is a function to be specified later. Then, using the bound (12), (16) and (17) give

$$\begin{aligned} \mathbb{E}\|\hat{f}_{\hat{m}} - f\|^2 &\leq \|f_m - f\|^2 + \frac{1}{5}\mathbb{E}\|\hat{f}_{\hat{m}} - f_m\|^2 + 160r_0^2 \frac{\mathbb{E}(\tau^2) + \mu(A)\mathbb{E}_A(\tau^4)}{n} \\ &\quad + 20 \sum_{m' \in \mathcal{M}_n} \mathbb{E}\left[ \sup_{t \in B(m')} \nu_n^{(3)}(t)^2 - p(m, m') \right] \\ &\quad + 20p(m, \hat{m}) + \text{pen}(m) - \text{pen}(\hat{m}). \end{aligned}$$

We choose  $p(m, m')$  such that  $20p(m, m') \leq \text{pen}(m) + \text{pen}(m')$ . Thus  $20p(m, \hat{m}) + \text{pen}(m) - \text{pen}(\hat{m}) \leq 2\text{pen}(m)$ . Let

$$Z_n(t) = \frac{1}{n} \sum_{j=1+\tau(1)}^{\tau(l_n)} [t(X_j) - \langle t, f \rangle]$$

and  $W(m, m') = [\sup_{t \in B(m')} Z_n^2(t) - p(m, m')]_+$ .

We use now the inequality  $\frac{1}{5}(x+y)^2 \leq \frac{1}{3}x^2 + \frac{1}{2}y^2$  to deduce

$$\begin{aligned} \mathbb{E}\|\hat{f}_{\hat{m}} - f\|^2 &\leq \frac{1}{3}\mathbb{E}\|\hat{f}_{\hat{m}} - f\|^2 + \frac{3}{2}\|f_m - f\|^2 + 20 \sum_{m' \in \mathcal{M}_n} \mathbb{E}W(m, m') + 2\text{pen}(m) + \frac{C}{n} \\ \Rightarrow \mathbb{E}\|\hat{f}_{\hat{m}} - f\|^2 &\leq \frac{9}{4}\|f_m - f\|^2 + 30 \sum_{m' \in \mathcal{M}_n} \mathbb{E}W(m, m') + 3\text{pen}(m) + \frac{3C}{2n}. \end{aligned}$$

We need now to bound  $\mathbb{E}W(m, m')$  to complete the proof. Proposition 2 below implies

$$\mathbb{E}W(m, m') \leq K' e^{-D_{m'}} (r_0 \vee 1)^2 K_3 \frac{1 + K_2 \|f\|_\infty}{n}$$

where  $K'$  is a numerical constant and  $K_2, K_3$  depend on the chain and with

$$(18) \quad p(m, m') = K \frac{D(m')}{n} (r_0 \vee 1)^2 K_3 (1 + K_2 \|f\|_\infty).$$

The notation  $a \vee b$  means  $\max(a, b)$ .

Assumption M3 yields  $\sum_{m' \in \mathcal{M}_n} e^{-D_{m'}} \leq \sum_{k \geq 1} e^{-k} = 1/(e-1)$ . Thus, by summation on  $m'$  in  $\mathcal{M}_n$

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}W(m, m') \leq K' \frac{1}{e-1} (r_0 \vee 1)^2 K_3 \frac{1 + K_2 \|f\|_\infty}{n}.$$

It remains to specify the penalty, which has to satisfy  $20p(m, m') \leq \text{pen}(m) + \text{pen}(m')$ . The value of  $p(m, m')$  is given by (18), so we set

$$\text{pen}(m) = 20K \frac{D_m}{n} (r_0 \vee 1)^2 K_3 (1 + K_2 \|f\|_\infty)$$

Finally

$$\forall m \quad \mathbb{E}\|\hat{f}_{\hat{m}} - f\|^2 \leq 3\|f_m - f\|^2 + 3\text{pen}(m) + \frac{C_1}{n}$$

where  $C_1$  depends on  $r_0, \|f\|_\infty, \mu(A), \mathbb{E}_\mu(\tau^2), \mathbb{E}_A(\tau^4), K_2, K_3$ . Since it is true for all  $m$ , we obtain the result.

*Second step:* We do not suppose any more that  $(X_n)$  has an atom.

The Nummelin splitting technique allows us to create the chain  $(X_n^*)$  and to define  $\gamma_n^*(t)$  and  $\hat{f}_m^*$  as above by (14),(15). Set now

$$\hat{m}^* = \arg \min_{m \in \mathcal{M}_n} [\gamma_n^*(\hat{f}_m^*) + \text{pen}(m)]$$

and  $\tilde{f}^* = \hat{f}_{\hat{m}^*}^*$ . The property P1 in section 5.1 gives  $\mathbb{E}\|f - \tilde{f}\|^2 = \mathbb{E}\|f - \tilde{f}^*\|^2$ . The split chain having an atom, we can use the first step to deduce

$$\mathbb{E}\|f - \tilde{f}^*\|^2 \leq 3 \inf_{m \in \mathcal{M}_n} \{d^2(f, S_m) + \text{pen}(m)\} + \frac{C_1}{n}.$$

And then the result is valid when replacing  $\tilde{f}^*$  by  $\tilde{f}$ . □

**Proposition 2.** *Let  $(X_n)$  be a Markov chain which satisfies assumptions A1–A5 and  $(S_m)_{m \in \mathcal{M}_n}$  be a collection of models satisfying the assumptions M1–M3. We suppose that  $(X_n)$  has an atom  $A$ . Let  $Z_n(t) = \frac{1}{n} \sum_{i=1+\tau(1)}^{\tau(l_n)} [t(X_i) - \langle t, f \rangle]$  where  $\tau(1)$  is the first return time in  $A$  and  $\tau(l_n)$  is the last return time. Let  $B = \{t \in S_m + S'_m, \|t\| = 1\}$  and*

$$W(m, m') = [ \sup_{t \in B} Z_n^2(t) - p(m, m') ]_+$$

where

$$p(m, m') = K \frac{D(m')}{n} (r_0 \vee 1)^2 \frac{1 + \|f\|_\infty \mathbb{E}_A(s^\tau)}{(\ln s)^2}$$

(where  $K$  is a numerical constant and  $s$  is a real depending on the chain). Then

$$\mathbb{E}W(m, m') \leq K' e^{-D_{m'}} (r_0 \vee 1)^2 \frac{1 + \|f\|_\infty \mathbb{E}_A(s^\tau)}{(\ln s)^2 n}$$

**Proof of Proposition 2:** We can write

$$Z_n(t) = \frac{1}{n} \sum_{j=1}^{l_n-1} S_j(t)$$

where  $S_j(t)$  is defined by (13). According to Lemma 1:

$\mathbb{E}_\mu |S_j(t)|^m \leq (2\|t\|_\infty)^{m-2} \|f\|_\infty \|t\|^2 \mathbb{E}_A(\tau^m)$ . Now, we use condition A5 of geometric ergodicity. The proof of Theorem 15.4.2 in Meyn and Tweedie (1993) shows that  $A$  is

a Kendall set, i.e. there exists  $s > 1$  (depending on  $A$ ) such that  $\sup_{x \in A} \mathbb{E}_x(s^\tau) < \infty$ . Then  $\mathbb{E}_A(\tau^m) \leq [m! / (\ln s)^m] \mathbb{E}_A(s^\tau)$ . Indeed

$$\begin{aligned} \mathbb{E}_A(\tau^m) &= \int_0^\infty m x^{m-1} P_A(\tau > x) dx \\ &\leq \int_0^\infty m x^{m-1} s^{-x} \mathbb{E}_A(s^\tau) dx = \frac{m!}{(\ln s)^m} \mathbb{E}_A(s^\tau) \end{aligned}$$

Thus

$$(19) \quad \forall m \geq 2 \quad \mathbb{E}_\mu |S_j(t)|^m \leq m! \left( \frac{2\|t\|_\infty}{\ln s} \right)^{m-2} \frac{\|f\|_\infty \|t\|^2}{(\ln s)^2} \mathbb{E}_A(s^\tau).$$

We use now the following inequality (see Petrov (1975) p.49) :

$$P\left(\max_{1 \leq l \leq n} \sum_{j=1}^l S_j(t) \geq y\right) \leq 2P\left(\sum_{j=1}^n S_j(t) \geq y - \sqrt{2B_n}\right)$$

where  $B_n \geq \sum_{j=1}^n \mathbb{E} S_j(t)^2$ . The inequality (19) gives us  $B_n = 2n \frac{\|f\|_\infty \|t\|^2}{(\ln s)^2} \mathbb{E}_A(s^\tau)$  and

$$P\left(\sum_{j=1}^{l_n-1} S_j(t) \geq y\right) \leq P\left(\max_{1 \leq l \leq n} \sum_{j=1}^l S_j(t) \geq y\right) \leq 2P\left(\sum_{j=1}^n S_j(t) \geq y - 2\sqrt{n}\|t\|M/\ln s\right)$$

where  $M^2 = \|f\|_\infty \mathbb{E}_A(s^\tau)$ . We use then the Bernstein inequality given by Birgé and Massart (1998).

$$P\left(\sum_{j=1}^n S_j(t) \geq n\varepsilon\right) \leq e^{-n\varepsilon}$$

with  $\varepsilon = \frac{2\|t\|_\infty}{\ln s}x + \frac{2\|t\|M}{\ln s}\sqrt{x}$ . Indeed, according to (19),

$$\frac{1}{n} \sum_{j=1}^n \mathbb{E} |S_j(t)|^m \leq \frac{m!}{2} \left( \frac{2\|t\|_\infty}{\ln s} \right)^{m-2} \left( \frac{\sqrt{2}\|t\|M}{\ln s} \right)^2.$$

Finally

$$(20) \quad P\left(Z_n(t) \geq \frac{2}{\ln s} \left[ \|t\|_\infty x + M\|t\|\sqrt{x} + M\|t\|/\sqrt{n} \right] \right) \leq 2e^{-n\varepsilon}.$$

We will now use a chaining technique used in Barron et al. (1999). Let us recall first the following lemma (Lemma 9 p.400 in Barron et al. (1999), see also Proposition 1 in Birgé and Massart (1998)).

**Lemma 2.** *Let  $\bar{S}$  a subspace of  $L^2$  with dimension  $D$  spanned by  $(\varphi_\lambda)_{\lambda \in \Lambda}$  (orthonormal basis). Let*

$$r = \frac{1}{\sqrt{D}} \sup_{\beta \neq 0} \frac{\|\sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda\|_\infty}{\sup_{\lambda \in \Lambda} |\beta_\lambda|}.$$

*Then, for all  $\delta > 0$ , we can find a countable set  $T \subset \bar{S}$  and a mapping  $\pi$  from  $\bar{S}$  to  $T$  such that :*

- *for all ball  $\mathcal{B}$  with radius  $\sigma \geq 5\delta$*

$$(21) \quad |T \cap \mathcal{B}| \leq (5\sigma/\delta)^D$$

- *$\|u - \pi(u)\| \leq \delta, \forall u \in \bar{S}$  et*

$$\sup_{u \in \pi^{-1}(t)} \|u - t\|_\infty \leq r\delta, \forall t \in T.$$

We apply this lemma to the subspace  $S_m + S_{m'}$  with dimension  $D_m \vee D_{m'}$  denoted by  $D(m')$  and  $r = r(m')$  defined by

$$r(m') = \frac{1}{\sqrt{D(m')}} \sup_{\beta \neq 0} \frac{\|\sum_{\lambda \in \Lambda(m')} \beta_\lambda \varphi_\lambda\|_\infty}{\sup_{\lambda \in \Lambda(m')} |\beta_\lambda|}$$

where  $(\varphi_\lambda)_{\lambda \in \Lambda(m')}$  is an orthonormal basis of  $S_m + S_{m'}$ . Notice that this quantity satisfy  $\phi_{m''} \leq r(m') \leq \sqrt{D(m')} \phi_{m''}$  where  $m''$  is such that  $S_m + S_{m'} = S_{m''}$  and then, using M2,

$$r(m') \leq r_0 \sqrt{D(m')}.$$

We consider  $\delta_0 \leq 1/5$ ,  $\delta_k = \delta_0 2^{-k}$ , and the  $T_k = T \cap B$  where  $T$  is defined by Lemma 2 with  $\delta = \delta_k$ . Inequality (21) gives us  $|T \cap B| \leq (5/\delta_k)^{D(m')}$  where  $B$  is the unit ball of  $S_m + S_{m'}$ . By letting  $H_k = \ln(|T_k|)$ , we obtain

$$(22) \quad H_k \leq D(m') \left[ \ln\left(\frac{5}{\delta_0}\right) + k \ln 2 \right].$$

Thus, for all  $u$  in  $B$ , we can find a sequence  $\{u_k\}_{k \geq 0}$  with  $u_k \in T_k$  such that  $\|u - u_k\| \leq \delta_k$  and  $\|u - u_k\|_\infty \leq r(m') \delta_k$ . Hence, we have the following decomposition:

$$u = u_0 + \sum_{k=1}^{\infty} (u_k - u_{k-1})$$

with  $\|u_0\| \leq 1$ ,  
 $\|u_0\|_\infty \leq r_0 \sqrt{D(m')} \|u_0\| \leq r_0 \sqrt{D(m')}$   
and for all  $k \geq 1$ ,

$$\begin{aligned} \|u_k - u_{k-1}\| &\leq \delta_k + \delta_{k-1} = 3\delta_{k-1}/2, \\ \|u_k - u_{k-1}\|_\infty &\leq 3r(m')\delta_{k-1}/2 \leq 3r_0 \sqrt{D(m')} \delta_{k-1}/2. \end{aligned}$$

Then

$$\begin{aligned} P(\sup_{u \in B} Z_n(u) > \eta) &= P(\exists (u_k)_{k \geq 0} \in \prod_{k \geq 0} T_k, Z_n(u_0) + \sum_{k=1}^{\infty} Z_n(u_k - u_{k-1}) > \eta_0 + \sum_{k=1}^{\infty} \eta_k) \\ &\leq \sum_{u_0 \in T_0} P(Z_n(u_0) > \eta_0) + \sum_{k=1}^{\infty} \sum_{\substack{u_k \in T_k \\ u_{k-1} \in T_{k-1}}} P(Z_n(u_k - u_{k-1}) > \eta_k) \end{aligned}$$

with  $\eta_0 + \sum_{k=1}^{\infty} \eta_k \leq \eta$ . We use the exponential inequality (20) to obtain

$$\begin{aligned} \sum_{u_0 \in T_0} P(Z_n(u_0) > \eta_0) &\leq 2e^{H_0 - nx_0} \\ \sum_{\substack{u_k \in T_k \\ u_{k-1} \in T_{k-1}}} P(Z_n(u_k - u_{k-1}) > \eta_k) &\leq 2e^{H_k + H_{k-1} - nx_k} \end{aligned}$$

by choosing  $\begin{cases} \eta_0 = \frac{2}{\ln s} \left( r_0 \sqrt{D(m')} x_0 + M \sqrt{x_0} + \frac{M}{\sqrt{n}} \right) \\ \eta_k = \frac{3}{\ln s} \left( r_0 \sqrt{D(m')} \delta_{k-1} x_k + M \delta_{k-1} \sqrt{x_k} + \frac{M \delta_{k-1}}{\sqrt{n}} \right). \end{cases}$

Let us choose now the  $(x_k)_{k \geq 0}$  such that  $nx_0 = H_0 + D_{m'} + v$  and for  $k \geq 1$ ,

$$nx_k = H_{k-1} + H_k + kD_{m'} + D_{m'} + v$$

Thus

$$P(\sup_{u \in B} Z_n(u) > \eta) \leq 2e^{-D_{m'} - v} \left( 1 + \sum_{k \geq 1} e^{-kD_{m'}} \right) \leq 3.2e^{-D_{m'} - v}$$

It remains to bound  $\sum_{k=0}^{\infty} \eta_k$ :

$$\sum_{k=0}^{\infty} \eta_k \leq \frac{1}{(\ln s)} (A_1 + A_2 + A_3).$$

$$\text{where } \begin{cases} A_1 &= r_0 \sqrt{D(m')} (2x_0 + 3 \sum_{k=1}^{\infty} \delta_{k-1} x_k) \\ A_2 &= 2M \sqrt{x_0} + 3M \sum_{k=1}^{\infty} \delta_{k-1} \sqrt{x_k} \\ A_3 &= 2 \frac{M}{\sqrt{n}} + \sum_{k=1}^{\infty} \frac{3M \delta_{k-1}}{\sqrt{n}} \end{cases}$$

- Regarding the third term, just write

$$A_3 = \frac{M}{\sqrt{n}} \left( 2 + 3 \sum_{k=1}^{\infty} \delta_{k-1} \right) = \frac{M}{\sqrt{n}} (6\delta_0 + 2) \leq c_5(\delta_0) \frac{M}{\sqrt{n}}$$

with  $c_5(\delta_0) = 6\delta_0 + 2$ .

- Let us bound the first term. First, recall that  $D(m') \leq \sqrt{n}$  and then

$$A_1 \leq r_0 \sqrt{\frac{n}{D(m')}} \left( 2 \frac{H_0 + D_{m'} + v}{n} + 3 \sum_{k=1}^{\infty} \delta_{k-1} \frac{H_{k-1} + H_k + kD_{m'} + D_{m'} + v}{n} \right).$$

Observing that  $\sum_{k=1}^{\infty} \delta_{k-1} = 2\delta_0$  and  $\sum_{k=1}^{\infty} k\delta_{k-1} = 4\delta_0$  and using (22), we get

$$A_1 \leq c_1(\delta_0) r_0 \frac{v}{\sqrt{n} D(m')} + 2c_2(\delta_0) r_0 \sqrt{\frac{D(m')}{n}}$$

$$\text{with } \begin{cases} c_1(\delta_0) &= 2 + 6\delta_0 \\ c_2(\delta_0) &= c_1(\delta_0) + \ln\left(\frac{5}{\delta_0}\right)(2 + 12\delta_0) + 6\delta_0(2 + 3 \ln 2) \end{cases}$$

- To bound the second term, we use the Schwarz inequality and the inequality  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ . We obtain

$$A_2 \leq c_3(\delta_0) M \sqrt{\frac{v}{n}} + c_4(\delta_0) M \sqrt{\frac{D(m')}{n}}$$

$$\text{with } \begin{cases} c_3(\delta_0) &= 6\delta_0 + 2 \\ c_4(\delta_0) &= 2\sqrt{1 + \ln\left(\frac{5}{\delta_0}\right)} + 3\sqrt{2\delta_0} \sqrt{(6\delta_0(1 + \ln 2) + 4\delta_0 \ln\left(\frac{5}{\delta_0}\right))} \end{cases}$$

We get so

$$\begin{aligned} \left(\sum_{k=0}^{\infty} \eta_k\right) &\leq \left(\frac{r_0 \vee 1}{\ln s}\right) \left(c_1 \frac{v}{\sqrt{nD(m')}} + c_3 M \sqrt{\frac{v}{n}}\right) \\ &\quad + \sqrt{\frac{D(m')}{n}} \left(\frac{r_0 \vee 1}{\ln s}\right) [c_2 + c_4 M + c_5 M] \\ &\leq c_6(\delta_0) \left(\frac{r_0 \vee 1}{\ln s}\right)^2 \left[\frac{v^2}{nD(m')} \vee M^2 \frac{v}{n}\right] + c_7(\delta_0) \frac{D(m')}{n} \left(\frac{r_0 \vee 1}{\ln s}\right)^2 (1+M)^2 \end{aligned}$$

$$\text{where } \begin{cases} c_6(\delta_0) = 6(c_1 + c_3)^2 \\ c_7(\delta_0) = \frac{6}{5} \sup(c_2, c_4 + c_5)^2 \end{cases}$$

Let us choose now  $\delta_0 = 0.024$  and then  $c_6 = 110$ ,  $c_7 = 268$ . Let  $K_1 = c_6(r_0 \vee 1 / \ln s)^2$ .

Then

$$\eta^2 = K_1 \left[\frac{v^2}{nD(m')} \vee M^2 \frac{v}{n}\right] + p(m, m')$$

where

$$p(m, m') = 2c_7(r_0 \vee 1)^2 \frac{D(m')}{n} \frac{1 + \|f\|_{\infty} \mathbb{E}_A(s^{\tau})}{(\ln s)^2}$$

$$\begin{aligned} \text{We get } &P(\sup_{u \in B} Z_n^2(u) > K_1 \left[\frac{v^2}{nD(m')} \vee M^2 \frac{v}{n}\right] + p(m, m')) \\ &= P(\sup_{u \in B} Z_n^2(u) > \eta^2) \\ &\leq P(\sup_{u \in B} Z_n(u) > \eta) + P(\sup_{u \in B} Z_n(u) < -\eta) \end{aligned}$$

Now

$$\begin{aligned} P(\sup_{u \in B} Z_n(u) < -\eta) &\leq \sum_{u_0 \in T_0} P(Z_n(u_0) < -\eta_0) + \sum_{k=1}^{\infty} \sum_{\substack{u_k \in T_k \\ u_{k-1} \in T_{k-1}}} P(Z_n(u_k - u_{k-1}) < -\eta_k) \\ &\leq \sum_{u_0 \in T_0} P(Z_n(-u_0) > \eta_0) + \sum_{k=1}^{\infty} \sum_{\substack{u_k \in T_k \\ u_{k-1} \in T_{k-1}}} P(Z_n(-u_k + u_{k-1}) > \eta_k) \\ &\leq 3.2e^{-D_{m'} - v}. \end{aligned}$$

Hence

$$P(\sup_{u \in B} Z_n^2(u) > K_1 \left[\frac{v^2}{nD(m')} \vee M^2 \frac{v}{n}\right] + p(m, m')) \leq 6.4e^{-D_{m'} - v}.$$

We obtain then

$$\begin{aligned}
\mathbb{E}[\sup_{t \in B} Z_n^2(t) - p(m, m')]_+ &\leq \int_0^\infty P(\sup_{u \in B} Z_n^2(u) > p(m, m') + z) dz \\
&\leq \int_0^{M^2 D(m')} P(\sup_{u \in B} Z_n^2(u) > p(m, m') + K_1 M^2 \frac{v}{n}) K_1 \frac{M^2}{n} dv \\
&+ \int_{M^2 D(m')}^\infty P(\sup_{u \in B} Z_n^2(u) > p(m, m') + K_1 \frac{v^2}{n D(m')}) K_1 \frac{2v}{n D(m')} dv \\
&\leq \frac{K_1}{n} \left[ M^2 \int_0^\infty 6.4 e^{-D_{m'} - v} dv + \frac{2}{D(m')} \int_0^\infty 6.4 e^{-D_{m'} - v} v dv \right] \\
&\leq \frac{6.4 K_1}{n} e^{-D_{m'}} \left( M^2 + \frac{2}{D(m')} \right) \\
&\leq 12.8 K_1 e^{-D_{m'}} \frac{1 + M^2}{n}.
\end{aligned}$$

By replacing  $M^2$  by its value, we get so

$$\mathbb{E}W(m, m') \leq K' \left( \frac{r_0 \vee 1}{\ln s} \right)^2 e^{-D_{m'}} \frac{1 + \|f\|_\infty \mathbb{E}_A(s^\tau)}{n}$$

where  $K'$  is a numerical constant □

**5.5. Proof of Corollary 2.** According to Theorem 1

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq C_2 \inf_{m \in \mathcal{M}_n} \left\{ d^2(f, S_m) + \frac{D_m}{n} \right\}$$

Since  $d^2(f, S_m) = O(D_m^{-2\alpha})$  (see Lemma 12 in Barron et al. (1999)),

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq C_3 \inf_{m \in \mathcal{M}_n} \left\{ D_m^{-2\alpha} + \frac{D_m}{n} \right\}$$

In particular, if  $m_0$  is such that  $D_{m_0} = \lfloor n^{\frac{1}{1+2\alpha}} \rfloor$ , then

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq C_3 \left\{ D_{m_0}^{-2\alpha} + \frac{D_{m_0}}{n} \right\} \leq C_4 n^{-\frac{2\alpha}{1+2\alpha}}.$$

The condition  $D_m \leq \sqrt{n}$  allows this choice of  $m$  only if  $\alpha > \frac{1}{2}$ . □

**5.6. Proof of Theorem 2.** The proof is identical to the one of Theorem 1. □

**5.7. Proof of Corollary 3.** It is sufficient to prove that  $d(g, S_m^{(2)}) \leq D_m^{-\alpha}$  if  $g$  belongs to  $B_{2,\infty}^\alpha([0, 1]^2)$ . It is done in the following lemma. □

**Lemma 3.** *Let  $g$  in the Besov space  $B_{2,\infty}^\alpha([0, 1]^2)$ . We consider the following spaces of dimension  $D^2$  :*

- $S_1$  is a space of piecewise polynomials of degree bounded by  $s > \alpha - 1$  based on a partition with square of vertice  $1/D$ ,

- $S_2$  is a space of orthonormal wavelets of regularity  $s > \alpha - 1$ ,
- $S_3$  is the space of trigonometric polynomials.

Then, there exists positive constants  $C_i$  such that

$$d(g, S_i) \leq C_i D^{-\alpha} \quad \text{for } i = 1, 2, 3.$$

*Proof of Lemme 3:* Let us recall the definition of  $B_{2,\infty}^\alpha([0, 1]^2)$ . Let

$$\Delta_h^r g(x, y) = \sum_{k=0}^r (-1)^{r-k} \binom{r}{k} g(x + kh_1, y + kh_2)$$

the  $r$ th difference operator with step  $h$  and

$$\omega_r(g, t) = \sup_{|h| \leq t} \|\Delta_h^r g\|_2$$

the  $r$ th modulus of smoothness of  $g$ . We say  $g$  is in the Besov space  $B_{2,\infty}^\alpha([0, 1]^2)$  if

$$\sup_{t>0} t^{-\alpha} \omega_r(g, t) < \infty$$

for  $r = \lfloor \alpha \rfloor + 1$ , or equivalently, for  $r$  an integer larger than  $\alpha$ .

DeVore (1998) proved that  $d(g, S_1) \leq C \omega_{s+1}(g, D^{-1})$ , so

$$d(g, S_1) \leq C D^{-\alpha}.$$

For the wavelets case, we use the fact that  $f$  belongs to  $B_{2,\infty}^\alpha([0, 1]^2)$  if and only if

$$\sup_{j \geq -1} 2^{j\alpha} \|\beta_j\| < \infty$$

(see Meyer (1990) chapter 6, section 10). If  $g_D$  is the orthogonal projection of  $g$  on  $S_2$ , it follows from Bernstein's inequality that

$$\|g - g_D\|^2 = \sum_{j>m} \sum_{k,l} |\beta_{jkl}|^2 \leq C \sum_{j>m} 2^{-2j\alpha} \leq C' D^{-j\alpha}$$

where  $m$  is such that  $2^m = D$ . For the trigonometric case, we will just adjust the demonstration of DeVore and Lorentz (1993) to the dimension 2. Let us begin to define  $F(x, y) = g(\frac{x}{2\pi}, \frac{y}{2\pi})$ . We will prove that

$$d(F, S_3(0, 2\pi)) \leq C \omega_r(F, D^{-1}) \quad \text{for } r = \lfloor \alpha \rfloor + 1,$$

then

$$d(g, S_3) \leq C \omega_r(g, D^{-1}) \leq C' D^{-\alpha}$$

Let  $K(t) = \lambda \left( \frac{\sin mt/2}{\sin t/2} \right)^{2r}$  the generalized Jackson Kernel where  $m = \lfloor D/r \rfloor + 1$  and  $\lambda$  is such that  $\int_0^{2\pi} K(t) = 1$ . We define now

$$SF(x, y) = \int [(-1)^{r+1} \Delta_{(t,u)}^r F(x, y) + F(x, y)] K(t) K(u) dt du$$

Let us notice that  $SF$  belongs to  $S_3$  since it is a linear combination of terms

$$\int F(x + kt, y + ku) \cos(lt) \cos(l'u) dt du$$

with  $k = 1, \dots, r$  and  $l, l' = 1, \dots, D$ . But  $t \mapsto F(x + kt, y + ku)$  is  $2\pi/k$  periodic, so the integral is zero unless  $k$  divides  $l$  and  $k$  divides  $l'$ . In the latter case, it is a trigonometric polynomial of degree  $l/k$  in  $x$  and  $l'/k$  in  $y$ . Thus  $SF$  is a trigonometric polynomial of degree  $\leq D$  and  $d(F, S_3) \leq \|SF - F\|$ .

$$\begin{aligned} \|SF - F\|_2 &\leq \left\| \int (-1)^{r+1} \Delta_{(t,u)}^r F K(t) K(u) dt du \right\| \\ &\leq \int \|\Delta_{(t,u)}^r F\| K(t) K(u) dt du \\ &\leq \int \omega_r(F, |t| \vee |u|) K(t) K(u) dt du \end{aligned}$$

But the  $r$ th modulus of smoothness satisfies the following property  $\omega_r(F, t) \leq (Dt + 1)^r \omega_r(F, D^{-1})$ , so

$$\begin{aligned} d(F, S_3) &\leq \omega_r(F, D^{-1}) \int (D(|t| \vee |u|) + 1)^r K(t) K(u) dt du \\ &\leq \omega_r(F, D^{-1}) \int (D|t| + 1)^r (D|u| + 1)^r K(t) K(u) dt du \\ &\leq \omega_r(F, D^{-1}) \left( \int_0^\pi (D|t| + 1)^r K(t) dt \right)^2 \end{aligned}$$

Now, Lemma 2.1 chapter 7 of DeVore and Lorentz (1993) shows that

$$\int_0^\pi (D|t| + 1)^r K(t) dt \leq C_r$$

and then  $d(F, S_3) \leq C \omega_r(F, D^{-1})$ . □

5.8. **Proof of Theorem 3.** Let us prove first the first item Let  $E_n = \{\|f - \tilde{f}\|_\infty \leq \frac{\chi}{2}\}$ . On  $E_n$ ,  $\tilde{f}(x) = \tilde{f}(x) - f(x) + f(x) \geq \frac{\chi}{2}$  and for  $n$  large enough,  $\tilde{\pi}(x, y) = \frac{\tilde{g}(x, y)}{\tilde{f}(x)}$ . For all  $(x, y) \in [0, 1]^2$ ,

$$\begin{aligned} |\tilde{\pi}(x, y) - \pi(x, y)|^2 &\leq \left| \frac{\tilde{g}(x, y) - \tilde{f}(x)\pi(x, y)}{\tilde{f}(x)} \right|^2 \mathbf{1}_{E_n} + (\|\tilde{\pi}\|_\infty + \|\pi\|_\infty)^2 \mathbf{1}_{E_n^C} \\ &\leq \frac{|\tilde{g}(x, y) - g(x, y) + \pi(x, y)(f(x) - \tilde{f}(x))|^2}{\chi^2/4} \\ &\quad + (a_n + \|\pi\|_\infty)^2 \mathbf{1}_{E_n^C} \\ \mathbb{E}\|\pi - \tilde{\pi}\|^2 &\leq \frac{8}{\chi^2} [\mathbb{E}\|g - \tilde{g}\|^2 + \|\pi\|_\infty^2 \mathbb{E}\|f - \tilde{f}\|^2] + (a_n + \|\pi\|_\infty)^2 P(E_n^C) \end{aligned}$$

It remains to bound  $P(E_n^C)$ .

$$\|f - \tilde{f}\|_\infty \leq \|f - f_{\hat{m}}\|_\infty + \|f_{\hat{m}} - \tilde{f}_{\hat{m}}\|_\infty$$

Let  $\gamma = \alpha - \frac{1}{2}$ , then  $B_{2,\infty}^\alpha([0, 1]) \subset B_{\infty,\infty}^\gamma([0, 1])$  (see DeVore and Lorentz (1993) p.182). Thus  $f$  belongs to  $B_{\infty,\infty}^\gamma([0, 1])$  and Lemma 12 in Barron et al. (1999) gives

$$\|f - f_{\hat{m}}\|_\infty \leq D_{\hat{m}}^{-\gamma} \leq k_n^{-\gamma}$$

Thus  $\|f - f_{\hat{m}}\|_\infty$  decreases to 0 and  $\|f - f_{\hat{m}}\|_\infty \leq \frac{\chi}{4}$  for  $n$  large enough. And so,

$$P(E_n^C) \leq P(\|f_{\hat{m}} - \tilde{f}_{\hat{m}}\|_\infty > \frac{\chi}{4})$$

But  $\|f_{\hat{m}} - \tilde{f}_{\hat{m}}\|_\infty \leq r_0 \sqrt{D(\hat{m})} \|f_{\hat{m}} - \hat{f}_{\hat{m}}\| \leq r_0 n^{1/8} \|f_{\hat{m}} - \hat{f}_{\hat{m}}\|$  and  $\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|^2 = \sum_{\lambda \in \Lambda(\hat{m})} \nu_n^2(\varphi_\lambda)$ . Thus,

$$\begin{aligned} P(E_n^C) &\leq P\left(\sum_{\lambda \in \Lambda(\hat{m})} \nu_n^2(\varphi_\lambda) > \frac{\chi^2}{16r_0^2 n^{1/4}}\right) \\ &\leq P\left(\sum_{\lambda \in \Lambda(\hat{m})} \nu_n^{(1)}(\varphi_\lambda)^2 + \nu_n^{(2)}(\varphi_\lambda)^2 + \nu_n^{(4)}(\varphi_\lambda)^2 > \frac{\chi^2}{32r_0^2 n^{1/4}}\right) \\ &\quad + P\left(\sum_{\lambda \in \Lambda(\hat{m})} Z_n^2(\varphi_\lambda) > \frac{\chi^2}{32r_0^2 n^{1/4}}\right) \\ &\leq \frac{32r_0^2 n^{1/4}}{\chi^2} \mathbb{E}\left(\sum_{\lambda \in \Lambda(\hat{m})} \nu_n^{(1)}(\varphi_\lambda)^2 + \nu_n^{(2)}(\varphi_\lambda)^2 + \nu_n^{(4)}(\varphi_\lambda)^2\right) \\ &\quad + \sup_{m \in \mathcal{M}_n} \sum_{\lambda \in \Lambda(m)} P(Z_n^2(\varphi_\lambda) > \frac{\chi^2}{32r_0^2 n^{1/4}}) \end{aligned}$$

We need then to bound two terms. For the first term, notice that  $|\nu_n^{(1)}(t)| \leq 2\|t\|_\infty \mathbf{1}_{\tau > n} \leq 2r_0 n^{1/8} \|t\| \mathbf{1}_{\tau > n}$  and so

$$\mathbb{E}(\nu_n^{(1)}(\varphi_\lambda)^2) \leq 4r_0^2 \frac{\mathbb{E}(\tau^2)n^{1/4}}{n^2}.$$

In the same way  $|\nu_n^{(2)}(t)| \leq 2\|t\| \frac{r_0 \tau n^{1/8}}{n}$  implies

$$\mathbb{E}(\nu_n^{(2)}(\varphi_\lambda)^2) \leq 4r_0^2 \frac{\mathbb{E}(\tau^2)n^{1/4}}{n^2}$$

At last, since  $|\nu_n^{(4)}(t)| \leq 2(n - \tau(l_n)) \frac{n^{1/8}}{n} r_0 \|t\| \mathbf{1}_{\tau \leq n}$

$$\mathbb{E}(\nu_n^{(4)}(\varphi_\lambda)^2) \leq 8r_0^2 \frac{\mu(A) \mathbb{E}_A(\tau^4)n^{1/4}}{n^2}$$

Thus

$$\mathbb{E}[\nu_n^{(1)}(\varphi_\lambda)^2 + \nu_n^{(2)}(\varphi_\lambda)^2 + \nu_n^{(4)}(\varphi_\lambda)^2] \leq Cn^{-7/4}$$

and

$$\mathbb{E}\left[\sum_{\lambda \in \Lambda(\hat{m})} \nu_n^{(1)}(\varphi_\lambda)^2 + \nu_n^{(2)}(\varphi_\lambda)^2 + \nu_n^{(4)}(\varphi_\lambda)^2\right] \leq \sup_{m \in \mathcal{M}_n} \sum_{\lambda \in \Lambda(m)} Cn^{-7/4} \leq Cn^{-3/2}$$

We can then bound the first term :

$$\frac{32r_0^2 n^{1/4}}{\chi^2} \mathbb{E}\left(\sum_{\lambda \in \Lambda(\hat{m})} \nu_n^{(1)}(\varphi_\lambda)^2 + \nu_n^{(2)}(\varphi_\lambda)^2 + \nu_n^{(4)}(\varphi_\lambda)^2\right) \leq C'n^{-5/4}$$

Besides, for all  $x$  and for all  $\lambda$ ,

$$P(Z_n(\varphi_\lambda) \geq 2r_0 n^{1/8} x + 2M\sqrt{x} + 2\frac{M}{\sqrt{n}}) \leq 2e^{-nx}$$

and so

$$P(Z_n^2(\varphi_\lambda) \geq (2r_0 n^{1/8} x + 2M\sqrt{x} + 2\frac{M}{\sqrt{n}})^2) \leq 4e^{-nx}$$

Let now  $x = n^{-1/2}$ ,  $x$  verifies (for  $n$  large enough)

$$2r_0 n^{1/4} x + 2Mn^{1/8} \sqrt{x} + 2Mn^{-3/8} \leq \frac{\chi}{r_0 \sqrt{32}}$$

that yields

$$(2r_0 n^{1/8} x + 2M\sqrt{x} + 2\frac{M}{\sqrt{n}})^2 \leq \frac{\chi^2}{32r_0^2 n^{1/4}}$$

The previous inequality gives then

$$P\left(Z_n^2(\varphi_\lambda) > \frac{\chi^2}{32r_0^2 n^{1/4}}\right) \leq 4e^{-nx} \leq 4e^{-\sqrt{n}}$$

Finally

$$P(E_n^C) \leq 4n^{1/4}e^{-\sqrt{n}} + C'n^{-5/4} \leq C''n^{-5/4}$$

for  $n$  great enough. And then, for  $n$  large enough,  $(a_n + \|\pi\|_\infty)^2 P(E_n^C) \leq Ca_n^2 n^{-5/4}$ . So, since  $a_n = o(n^{1/8})$ ,  $(a_n + \|\pi\|_\infty)^2 P(E_n^C) = o(\frac{1}{n})$ .

Following result in Theorem 3 is provided by using Corollary 2 and Corollary 3.

□

#### ACKNOWLEDGEMENTS

I would like to thank F. Comte for her helpful suggestions throughout this work.

#### REFERENCES

- A. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalization, *Probab. Theory Related Fields* **113** (1999), no. 3, 301–413
- A. K. Basu and D. K. Sahoo, On Berry-Esseen theorem for nonparametric density estimation in Markov sequences, *Bull. Inform. Cybernet.* **30** (1998), no. 1, 25–39
- L. Birgé and P. Massart, Minimum contrast estimators on sieves: exponential bounds and rates of convergence, *Bernoulli* **4** (1998), no. 3, 329–375
- D. Bosq, Sur l'estimation de la densité d'un processus stationnaire et mélangeant, *C. R. Acad. Sci. Paris Sér. A-B* **277** (1973), A535–A538
- S. J. M. Cléménçon, Méthodes d'ondelettes pour la statistique non paramétrique des chaînes de Markov, PhD thesis, Université Denis Diderot Paris 7, 1999
- S. J. M. Cléménçon, Adaptive estimation of the transition density of a regular Markov chain, *Math. Methods Statist.* **9** (2000), no. 4, 323–357
- S. J. M. Cléménçon, Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the Nummelin splitting technique, *Statist. Probab. Lett.* **55** (2001), no. 3, 227–238
- F. Comte and Y. Rozenholc, Adaptive estimation of mean and volatility functions in (auto-)regressive models, *Stochastic Process. Appl.* **97** (2002), no. 1, 111–145
- F. Comte and Y. Rozenholc, A new algorithm for fixed design regression and de-noising, *Ann. Inst. Statist. Math.* **56** (2004), no. 3, 449–473

- C. Dalelane, Data driven kernel choice in nonparametric density estimation, PhD thesis, Technische Universität Braunschweig, 2005; <http://opus.tu-bs.de/opus/volltexte/2005/659/>
- R. A. DeVore and G. G. Lorentz, *Constructive approximation*, Springer, Berlin, 1993
- R. A. DeVore, Nonlinear approximation, in *Acta numerica*, 51–150, Cambridge Univ. Press, 1998
- D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, Density estimation by wavelet thresholding, *Ann. Statist.* **24** (1996), no. 2, 508–539
- J. L. Doob, *Stochastic processes*, Wiley, New York, 1953
- P. Doukhan and M. Ghindès, Estimation de la transition de probabilité d’une chaîne de Markov Doëblin-récurrente. Étude du cas du processus autorégressif général d’ordre 1, *Stochastic Process. Appl.* **15** (1983), no. 3, 271–293
- P. Doukhan, *Mixing Properties and examples*, Springer, New York, 1994
- W. R. Gilks, S. Richardson and D. J. Spiegelhalter (editors) *Markov chain Monte Carlo in practice*, Chapman & Hall, London, 1996
- H. Gillert and A. Wartenberg, Density estimation for nonstationary Markov processes, *Math. Operationsforsch. Statist. Ser. Statist.* **15** (1984), no. 2, 263–275
- O. Hernández-Lerma, S. O. Esparza and B. S. Duran, Recursive nonparametric estimation of nonstationary Markov processes, *Bol. Soc. Mat. Mexicana (2)* **33** (1988), no. 2, 57–69
- R. Höpfner and E. Löcherbach, Limit theorems for null recurrent Markov processes, *Mem. Amer. Math. Soc.* **161** (2003), no. 768, vi+92 pp.
- I. A. Ibragimov and R. Z. Has’minskiĭ, An estimate of the density of a distribution, *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **98** (1980), 61–85, 161–162, 166
- S. F. Jarner and E. Hansen, Geometric ergodicity of Metropolis algorithms, *Stochastic Process. Appl.* **85** (2000), no. 2, 341–361
- F. Leblanc, Density estimation for a class of continuous time processes, *Math. Methods Statist.* **6** (1997), no. 2, 171–199
- E. Liescher, Density estimation for Markov chains, *Statistics* **23** (1992), no. 1, 27–48
- E. Masry and L. Györfi, Strong consistency and rates for recursive probability density estimators of stationary processes, *J. Multivariate Anal.* **22** (1987), no. 1,

79–93

- K. L. Mengersen and R. L. Tweedie, Rates of convergence of the Hastings and Metropolis algorithms, *Ann. Statist.* **24** (1996), no. 1, 101–121
- Y. Meyer, *Ondelettes et opérateurs. I*, Hermann, Paris, 1990
- S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer, London, 1993
- S. P. Meyn and R. L. Tweedie, Computable bounds for geometric convergence rates of Markov chains, *Ann. Appl. Probab.* **4** (1994), no. 4, 981–1011
- A. Mokkadem, Sur un modèle autorégressif non linéaire: ergodicité et ergodicité géométrique, *J. Time Ser. Anal.* **8** (1987), no. 2, 195–204
- E. Nummelin, *General irreducible Markov chains and nonnegative operators*, Cambridge Univ. Press, Cambridge, 1984
- V. V. Petrov, *Sums of independent random variables*, Translated from the Russian by A. A. Brown, Springer, New York, 1975
- B. L. S. Prakasa Rao, Density estimation for Markov processes using delta-sequences, *Ann. Inst. Statist. Math.* **30** (1978), no. 2, 321–328
- G. O. Roberts and J. S. Rosenthal, Markov-chain Monte Carlo: some practical implications of theoretical results, *Canad. J. Statist.* **26** (1998), no. 1, 5–31
- M. Rosenblatt, in *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, 199–213, Cambridge Univ. Press, London, 1970
- G. G. Roussas, Nonparametric estimation in Markov processes, *Ann. Inst. Statist. Math.* **21** (1969), 73–87
- T. Runst and W. Sickel, *Sobolev spaces of fractional order, Nemytskij operators, and nonlinear partial differential equations*, de Gruyter, Berlin, 1996
- K. Tribouley and G. Viennet,  $L_p$  adaptive density estimation in a  $\beta$  mixing framework, *Ann. Inst. H. Poincaré Probab. Statist.* **34** (1998), no. 2, 179–208
- S. Yakowitz, Nonparametric density and regression estimation for Markov sequences without mixing assumptions, *J. Multivariate Anal.* **30** (1989), no. 1, 124–136